Oral narrative tasks and second language performance : an investigation of task characteristics

Tavakoli, Parvaneh

Oral Narrative Tasks and Second Language Performance:

An Investigation of Task Characteristics,

~~Performance Conditions and~~

~~Test-takers' Perceptions of Task Difficulty~~

Parvaneh Tavakoli

Department of Education & Professional Studies

King's College, University of London

September 2004

A dissertation submitted in partial fulfilment of the requirements for the

Degree of Doctor of Philosophy of the University of London

ABSTRACT

At the centre of the debate on task-based language assessment is the issue of task difficulty and whether it is influenced by task characteristics and performance conditions. The overarching aim of the research reported here is to investigate the effects of characteristics and performance conditions of oral narrative tasks on second language performance in an assessment setting.

Study One sets out to investigate the impact of degree of task structure, pre-task planning and language proficiency levels on the language performance of 80 Iranian test-takers of English. It further attempts to explore the test-takers' perceptions of task difficulty. A quantitative methodological approach is taken and factorial designs for investigating between and within-participants differences are employed. Data are collected through recording the test-takers' performance on four oral narrative tasks and through retrospective questionnaires on task difficulty. The data are first transcribed and coded and eventually examined by employing a number of different statistical analyses.

Findings of this study have indicated that task structure directly influences accuracy and fluency of performance, i.e. performance on structured tasks is more accurate and fluent than performance on unstructured tasks. Moreover, performance on problem-solution structure has proved to be more fluent than performance on schematic sequential structure. However, the results suggest that task structure does not essentially influence complexity of performances on tasks. Results have further shown that pre-task planning has helped the test-takers improve different aspects of

their performance, i.e. fluency, accuracy and complexity. Interestingly, having the opportunity of pre-task planning, when compared to having a higher proficiency level, leads to greater fluency of performance on some fluency measures. Furthermore, the findings of this study reveal that the test-takers have predominantly perceived unstructured tasks as more difficult than structured tasks under both planned and unplanned conditions.

Based on the findings of Study One, Study Two is primarily designed to investigate the effects of grounding (foreground and background information) on language performance in general and on complexity of performance in particular. To confirm the findings of Study One and to examine the interactional effects of different task characteristics on language performance, task structure is further employed in Study Two. Data are collected through recording the language performance of 60 Iranian test-takers of English on six oral narrative tasks. Parallel to Study One, the data are examined through a number of statistical analyses.

Results of the second study confirm the findings of Study One regarding the effect of task structure. In addition, the results indicate that grounding influences the test-takers' performance in terms of its complexity and fluency. In effect, tasks which contain both foreground and background information have elicited language performance with greater syntactic complexity. A primary tradeoff relationship is observed between complexity and fluency of the performance on tasks. However, the relationship between accuracy and complexity is of a more delicate nature and is more likely to change as different task characteristics vary.

ACKNOWLEDGEMENTS

A number of people have contributed to making this thesis possible. First of all, I would like to thank my two supervisors, Professor Peter Skehan and Dr. Constant Leung, for their encouragement and support, both intellectual and moral. Their sensitive and thoughtful supervision has been of enormous benefit in extending my thinking and my understanding of research. I am greatly indebted to both of them, particularly for their unfailing support and extreme generosity in the amount of time they have provided me with.

I am especially grateful to Dr Pauline Foster for her invaluable contributions to this research. She always generously found the time to train me how to code the data, check the data transcripts, discuss my research, suggest new perspectives and generally respond to a range of queries.

I acknowledge with gratitude the help and advice Dr. Carys Jones has given me with the preparation of this thesis.

I extend my sincere thanks to Dr Hossein Farhady who taught me the principles of research and unsparingly supported me at every single stage of my post-graduate studies.

I would also like to thank the students, teachers and managers at Simin Educational Association, who offered their valuable assistance during the data collection.

Most of all, I should like to thank my family and would particularly like to offer my heart-felt thanks to my husband for his years of patience and help, without which my achievements would not have been possible.

Table of Contents

Chapter III
Task-Based Assessment

Chapter IV
Variables in Task-Based Research

Chapter XI
Conclusions, Implications and Suggestions for Further Research

CHAPTER I

Introduction

1.1 Background to This Research

Over the past two decades, studying characteristics and performance conditions of tasks has become a burgeoning area of research within task-based language teaching and assessment. Research on Second Language Acquisition (SLA) has shown that task characteristics and conditions directly influence task difficulty and second language learners' performance. Familiarity with the task and the topic, familiarity with the interlocutor(s), availability of pre-task planning time and complexity of the information provided in the task are some of the characteristics and performance conditions that influence task difficulty (Bygate, 2001; Crookes, 1989; Foster & Skehan, 1997).

Although research on language teaching has focused on different characteristics and performance conditions of tasks that would influence task difficulty and language performance in instructional settings, not many studies have been carried out to investigate the difficulty of tasks as units of language assessment. For a long time, tasks in language assessment were taken to be of equal difficulty (Fulcher & Reiter, 2003). It is only recently that language testing (LT) research acknowledges that tasks are of different difficulty levels and appreciates that a hierarchy of task difficulty has yet to be established (Bachman, 2002; Iwashita, McNamara, and Elder, 2001). LT research further

concedes that, in assessing oral ability, the oral ability continuum is not well defined because of the differences in performance outcomes resulting from task difficulty (Iwashita et al., 2001). Bachman (2002) argues that understanding the effects of tasks on language performance and how test-takers interact with tasks is " the most pressing issue facing language performance assessment" (p. 471). Hence, identifying task characteristics that determine task difficulty is now considered an important current challenge for LT researchers, since an index of task difficulty will be essential in selecting appropriate tasks, in designing and developing task-based assessment, in providing a more reliable assessment of oral ability and for the validity of the interpretations and uses that are made based on the test results.

1.2 Focus of the Research

Every year large numbers of people take language tests which would significantly influence their lives. Test results generally act as gateways at important transitional moments in their education, employment, and immigration. In practice, language tests are powerful decision making devices that affect an individual's life, particularly in the light of the fact that " it is the performance on a single test, often on one occasion at a single point in time, that can lead to irreversible, far-reaching and high-stake decisions" (Shohamy, 2001b, p. 16). Therefore, a significant goal for LT organizations is ensuring a fair assessment of test takers' language ability that is not influenced by ability-irrelevant factors.

In the context of testing English as a second language, many international tests (e.g. TOEFL's TSE and Cambridge's YLE) employ oral narrative tasks to elicit samples of

test-takers' language performance. This performance would then be assessed and used as a representation of the test-takers' oral language ability. However, SLA research has shown that there are certain characteristics and conditions of oral narrative tasks which would influence task difficulty and language performance on tasks. It would therefore seem important to take task difficulty into account when selecting tasks for the purpose of assessment. Furthermore, since performance on a task may be influenced by task difficulty, test results obtained from such performance may not be a true representation of a test-taker's language ability. Given that an index of task difficulty is not clearly defined yet, it would be very difficult to predict a test-taker's performance on other tasks. In addition, if task difficulty affects performance on tasks, this would in turn affect generalizability of the test results: "the extent to which our inferences generalize across a set of assessment tasks" (Bachman, 2002, p. 458). Evidently, systematic research is required to find out how task characteristics and conditions would influence task difficulty and second language performance on tasks.

1.3 Research Questions

The overarching aim of this research is to find out empirically whether there are certain characteristics and performance conditions of oral narrative tasks which would influence task difficulty, language performance on tasks and test-takers' perceptions of task difficulty. Thus in this research I will attempt to investigate whether:

- There are characteristics and performance conditions of oral narrative tasks that influence task difficulty in an assessment setting.

- There are characteristics and performance conditions of oral narrative tasks that influence test-takers' language performance on tasks.
- There are characteristics and performance conditions of oral narrative tasks that influence test-takers' perceptions of task difficulty.

It should be mentioned that these are the general research questions which will later turn into more specific questions in two interrelated studies.

1.4 Research Methodology

A quantitative approach to research is adopted in the current research. There are three prime reasons. First, as this research is being carried out in the context of task-based language assessment, it is very important to have conformity with the relevant literature from which the study has been drawn. Second, as SLA researchers (McMillan, 1996; Oppenheim, 1992) have emphasized, a quantitative approach is considered more appropriate for this type of research because both studies in the current research will deal with measurement and large numbers of participants. Third, adopting a quantitative approach appears to be more suitable for this research since the results of these studies are expected to be generalizable to similar populations and tasks. A limitation of this approach, however, is that the data would not permit investigating task difficulty from a qualitative approach. Undoubtedly, a qualitative study of task characteristics would be able to explore how an individual test-taker experiences the detailed aspects of task characteristics and performance conditions (For examples of qualitative studies on tasks see Dotano, 1994 and Duff, 1993).

1.5 Definition of Terms

A number of key terms are used in this research. A brief definition of each term is given here. However, detailed definitions and the relevant discussions will be presented later in chapters II, III and IV.

1) Task

There are several definitions of task in the language teaching and testing literature. Skehan (1996) has presented a definition of task, which is widely agreed upon, frequently referred to in the literature and is followed in this study. Task, as Skehan (1996a) defines, is " an activity in which meaning is primary; there is some sort of relationship to the real world; task completion has some priority; and the assessment of task performance is in terms of task outcome" (p.38).

2) Oral narrative Task

In the current context of language testing, 'oral narrative task' refers to stories based on a sequenced set of picture prompts which are given to test-takers in order to elicit samples of oral language performance. The term was primarily meant to include the whole process of performing the task, i.e. looking at the picture story, narrating the story to a listener and ensuring that the message gets through. However, the concept of 'oral narrative task', or 'narrative task', has been narrowed down and is now usually used to refer to the actual picture stories.

3) Task-Based Language Assessment

In task-based language assessment second language ability is attested through a number of tasks that are employed to elicit samples of a test-taker's language performance. The tasks are intended to " stimulate the language demands of the real world situation with the

aim of eliciting an 'authentic' sample of language from the candidate (Elder, Iwashita and McNamara, 2002, p. 347)".

4) Language Performance

Language performance is usually taken as a concept that is generally well understood so much so that there are no specific definitions provided in the literature to describe it. In this research, language performance refers to the oral language production of learners/test-takers when they employ their speaking ability to communicate meaning through language.

5) Fluency

Fluency, in the context of second language teaching and testing, generally refers to ease or automaticity in a learner's speech and represents flow, continuity and smoothness of speech. Skehan (1998) discusses fluency in terms of the learner's capacity to communicate meaning in real time, which reflects underlying speech-planning and thinking processes and shows the smoothness of oral performance. In the present research, Skehan's definition of fluency is adopted and a wide range of fluency measures, such as temporal and repair fluency measures, is employed to investigate the multifaceted nature of fluency.

6) Accuracy

The ability to avoid errors in performance is generally defined as accuracy. In this research, accuracy is particularly reflecting both higher levels of control in the language and avoidance of challenging structures that might provoke errors. In the current research, accuracy is represented through the ratio of error-free clauses, i.e. clauses that have no errors in their syntax, morphology, native-like lexical choice or word order.

7) Complexity

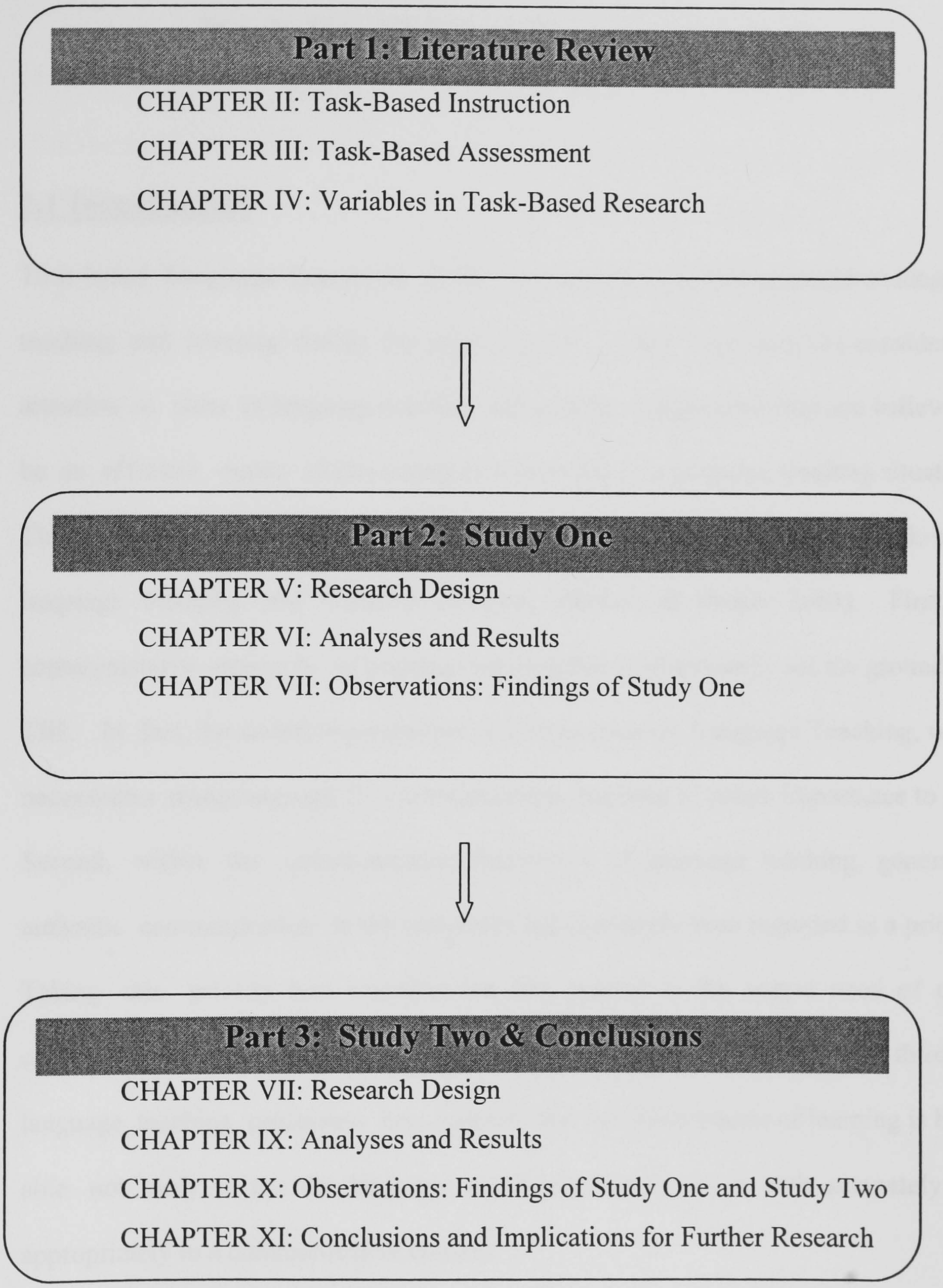
Complexity, or syntactic complexity, of performance is taken to refer to the level of organization of what a learner or test-taker says in terms of the variety of syntactic patterning and subordination he/she employs in his/her performance. Complexity, in other words, demonstrates the range of forms that surface in language production as well as the degree of sophistication of such forms. In the present research, complexity is measured through an index of subordination.

1.6 Overview of the Research

The current research reports on two interrelated studies and will be presented in three main parts . First, I will start with a review of the relevant literature on task-based instruction, task-based assessment and variables in task-based research. Then, I will focus on Study One, giving an account of the variables of the study, the research design, pilot study, main study, data, coding, measures, analyses, results and the relevant discussions and conclusions. Next, I will explain how Study Two is built on the results of Study One and will then focus on the research design of Study Two, pilot study, main study, data, coding, measures, analyses and results. Finally, I will discuss the conclusions to be drawn from the two studies reported here and will reflect on the implications these findings have and suggestions for further research. Every effort will be made to present the two studies in broadly the same way, with the exception of Figure 9.1, which has been added to provide the reader with a better access to the information being presented. Figure 1.1 is a flow chart of the structure of the current research.

Figure 1.1

Flow Chart: Structure of the Research



CHAPTER II

Task-Based Language Instruction

2.1 Introduction

Task-based Language Instruction (TBI) has become a popular approach to language teaching and learning during the past decades. Tasks have received considerable attention as units of language teaching and syllabus design since they are believed to be an effective means of generating communication in language teaching situations. Three significant themes are reported to be influencing the emergence of task-based language teaching and research (Bygate, Skehan, & Swain; 2001). First, the communicative approach to language teaching has fundamentally set the grounds for TBI. In fact, the underlying principle of Communicative Language Teaching, which necessitates using language for communication, has been of prime importance to TBI. Second, within the communicative framework of language teaching, generating authentic communication in the classroom has constantly been regarded as a priority. Taking this priority into consideration has pointed to the urgent need of using communicative tasks to meet classroom communication requirements. And third, the language teaching profession has realized that the effectiveness of learning is being able not only to use the language accurately but to use it both accurately and appropriately in a communicative context.

This chapter will first discuss the emergence and development of the Communicative Approach to second language teaching including the concepts and theoretical assumptions involved and the underlying pedagogic principles of the approach.

Discussions of 'Focus on Form' will be presented to set the grounds for moving from the communicative approaches to task-based approaches to language teaching. The chapter will then discuss the emergence of TBI, its principles in both pedagogy and research. The dominant research-oriented approaches to TBI will be then introduced and evaluated. From a conceptual perspective, "task", task difficulty, task characteristics and the role they have in language teaching will be defined in a following section of the current chapter. Finally, two salient proposals of task characteristics will be introduced and compared. It is worth noting that issues related to language assessment, particularly the task-based approach to oral language assessment, will be presented in the chapter that follows.

2.2 Communicative Approaches to Language Teaching

2.2.1 Background

Communication has generally been, for some time now, the recognized goal of most language teaching methods (Littlewood, 1999). However, for a period of time it was assumed that the route to this goal was through mastering the structures and vocabulary of the language. Traditionally, the most common approach to teaching a language was to present a structure, drill it, practice it in a context and move on to the next structure. The conventional pedagogic practice, therefore, was to move towards segregation, rather than integration, of language structures and rules. As a result, most second language courses consisted of units of comprehension, grammar and composition as separate sections.

This structure-based approach further influenced syllabus design of the materials written for language teaching. Syllabus design focused on selecting structural items and grading them in a suitable order for teaching. Language teaching syllabi were

often little more than ordered lists of structures. The mere task of syllabus design seemed to be building up the "synthetical" inventory of structural items which learners could handle (Wilkins, 1976).

During the 1970s a large number of second language teachers and researchers started shifting emphasis from teaching structures to teaching language as communication (Wilkins, 1976; Widdowson, 1978; Brumfit & Johnson, 1979). This shift mainly resulted from disorientation and disappointment of language teachers who employed structure-based approaches to language teaching. Researchers realized that this focus on structures would not necessarily lead learners to develop an ability to use language in real communication. Widdowson (1978) suggested that a distinction be made between language usage and language use. He defined the former as the citation of words and sentences as manifestation of the language system, and the latter as the way the system is realized for normal communicative purposes. It was then clear that although knowing a language was often taken to mean having proper knowledge of correct usage, this knowledge would be of little value on its own. This knowledge of language correctness needed to be complemented by knowledge of appropriateness in the context of communication.

Teaching language as communication seemed to be the trigger of a move towards an approach which would bring linguistic and communicative abilities into close association with each other. This led SLA researchers and language teachers to explore how they could focus on a more flexible framework that was adaptable to real use of language in communication. This move, which was supported by modern theories of linguistics, sociolinguistics and psycholinguistics, came to be widely known as the Communicative Language Teaching Approach (Brumfit & Johnson, 1979; Widdowson, 1978).

At the core of the Communicative Approach, there is a substantial amount of attention being paid to the nature of communication and to the role language plays in it. Achieving linguistic competence proved to be insufficient in learning a second language. What seemed indispensable was to go beyond the linguistic competence. However, one salient feature of the Communicative Approach was the distinction it made between the notions of "communicative" and "linguistic" competence. Since this distinction represents a significant principle of the Communicative Approach, it will be defined and discussed in the following section.

2.2.2 Communicative Competence

The term "communicative competence" was first used by Hymes (1972) to make a deliberate contrast to Chomsky's "linguistic competence". Chomsky (1965) used the term to refer to the speaker-hearer's knowledge of his/her language. Linguistic competence was, in fact, understood as the tacit knowledge of language structures available to an ideal speaker and/or listener and was distinguishes from performance which referred to the actual language behaviour. On the other hand, "communicative competence" was used to reflect the social view of language, which had found increasing acceptance since middle of the 1960s. Critical of Chomsky's account, Hymes (1972) argued that linguistic competence was only part of what a speaker knows in communicating with others. Besides mastering linguistic forms, Hymes discussed that speakers needed to know when, how and to whom it is appropriate to use these forms. He emphasized the prominent role of sociocultural factors in using language for communication. He proposed that communicative competence is what supposedly enables speakers to engage in socially appropriate exchanges.

In his further works, Hymes (1972) introduced four components of communicative competence:

> "I would suggest, then, that for language and other forms of communication (culture), four questions arise:
>
> 1. Whether (and to what degree) something is formally *possible*;
> 2. Whether (and to what degree) something is *feasible* in virtue of the means of implementation variable;
> 3. Whether (and to what degree) something is *appropriate* (adequate, happy, successful) in relation to a context in which it is used and evaluated;
> 4. Whether (and to what degree) something is in fact done, actually *performed*, and what its doing entails". (p. 281)

Following Hymes, other researchers reflected on the concept of communicative competence in the dominant research areas of the time. Canale and Swain (1980) defined communicative competence as the underlying systems of knowledge and skill required for communication. They proposed a three-component framework for communicative competence in 1980. Canale (1983) refined it to a four-component framework comprising grammatical competence, sociolinguistic competence, discourse competence and strategic competence. In their framework, Chomsky's linguistic competence was renamed as grammatical competence and referred to as 'mastery of linguistic code' (Canale, 1983). Grammatical competence, as defined by Canale and Swain (1980), included knowledge of lexical items, rules of morphology, syntax, sentence grammar and principles of semantics. Their concept of sociolinguistic competence was derived from Hymes' notion of appropriacy and understanding of the speakers in terms of the social relationship among the

participants. Discourse competence manifests speakers' knowledge of language beyond sentence level. It refers to the speaker's understanding of how written and spoken texts are organised and how meaning would be inferred and interpreted from the text by employing the organizational patterns of discourse. The last component, strategic competence, refers to the strategies that are called into action to compensate for the insufficient knowledge of the speaker or for the breakdowns that might happen in communication. Paraphrasing, repetition, hesitation, circumlocution, and avoidance are some of the strategies language speakers use to compensate for lack of knowledge or breakdowns in communication (Savignon, 1983).

Bachman (1990) and Bachman and Palmer (1996) have built upon Canale and Swain's model of communicative competence and proposed a new language ability model which is principally different from theirs. This model has made extensive contributions to the fields of second language acquisition and second language testing. Since it is more related to theories of language testing, Bachman and Palmer's model (1996) will be discussed in detail in the following chapter.

As well as the extended notion of communicative competence, the Communicative Approach draws on a number of other principles and assumptions. In the following section, significant principles of this approach will be introduced and discussed.

2.2.3 Principles of Communicative Language Teaching

With the introduction of the communicative approach in general and communicative competence in particular, and as a result of discussions raised in research and pedagogy, definitions of communication underwent critical reformulations. Studies of communicative competence revealed that there were complex relationships between structures and meanings as structures included conceptual meanings as well as the

communicative functions of language. It became clear that communication is essentially a process of interaction in which meanings are developed and negotiated over longer stretches of discourse (Wilkins, 1999). Therefore, it is necessary that learners develop skills to organize information and create links over longer stretches of writing, acquire techniques of opening and closing conversations, and learn techniques of agreeing or disagreeing with others.

In addition to the extended notions of genuine communication, the communicative approach has introduced a new perspective towards the concepts of learning (Littlewood, 1992; Lightbown & Spada, 1999). There have been traditional approaches to language learning in which learning has been considered as a form of skills developing. Items of language are taken as discrete, disintegrated parts being taught by the teachers and practiced by the learners until automatic mastery is achieved. Then learners would try to incorporate these items in to real-life communication. However, it is doubtful that the learned items of language would be integrated into a functioning system. In contrast, the communicative approach views learning as a natural growth. It is believed that activities that involve real communication promote learning. In addition, language that is meaningful to the learner as well as activities which are used for carrying out meaningful tasks will advance learning.

A distinctive feature of the communicative language teaching is the way syllabuses are defined and employed in this approach. Initially, Wilkins (1976) distinguished between *synthetic* and *analytic* syllabuses by arguing that synthetic syllabuses segment language into discrete linguistic items for presentation at a time. On the other hand, as Wilkins contended, analytic syllabuses "are organized in terms of the purposes for which people are learning language and the kinds of language

performance that are necessary to meet that purpose" (Wilkins, 1996, p.13). Since analytic syllabuses attempted to prioritize how the language is learnt, rather than what is to be learnt, and considered learners needs and purposes as the main criteria, they became a significant aspect of the communicative language teaching.

Research into interlanguage studies has demonstrated that learners have their own active mechanisms for receiving the input, constructing knowledge and developing their interlanguage system (Ellis, 1985; Tarone, 1979). Central to this is exposure to language and the motivation to use it for real communication. Within this framework, it is believed that a naturalistic exposure to language is a prerequisite to developing the interlanguage system. In fact, syntactic structures are assumed to develop through exposure to language and through the interactions learners have with speakers of the language and with one another (Hatch, 1978). A significant responsibility of language teachers, therefore, would be to provide learners with the opportunities in which language is processed through natural exposure and use. The learning processes by which the learner operates are natural and built in to the learner and cannot be simply determined by teachers (Skehan, 1996 b).

Another significant principle of communicative language teaching is its emphasis on learning by doing. Learners are constantly encouraged to do things with the language they are learning: the kinds of activities they recognize as purposeful communication; activities which resemble what they use their own language for (Widdowson, 1990). The basic "units" used by teachers to generate communication among learners were first called communicative activities (Skehan, 2003). Later, such learning activities that were organized around a topic or curriculum area and needed interaction with others were called "task" (Leung, Harris & Rampton, 2004).

Focus on meaning is another salient principle of communicative language teaching. It is believed that real-life communication occurs while participants prioritize meaning over forms of language. In the early days of the communicative approach, it was discussed that, having been exposed to language – comprehensible input, learners could work out the grammar for themselves (Krashen, 1980). The learner's language system would therefore automatically develop without language-focused instruction (Krashen & Terrell 1983). Formal instruction of form was discouraged, as it would lead up to learning rather than to acquisition[1] of a language, and all the attention had to be paid to meaning. However, this account of language teaching has been repeatedly criticized by different researchers. At a broad level, it is argued that Krashen's work has provided the language teaching literature with only some general principles of pedagogy which, in terms of the particular cases of language teaching such as tasks, are not of great significance (Richards, 1985). At a more specific level, the value of Krashen's view to formal instruction is often questioned. For instance, Pienemann (1984) argues that language instruction is highly beneficial when the learner is developmentally ready to acquire a particular linguistic feature. Furthermore, a number of research studies (Long, 1983, 1988; Doughty & William, 1998; Long & Robinson, 1998; Norris & Ortega, 2000) have shown that formal instruction has a positive effect on learning. Long (1983) and Long and Robinson (1998), among others, have argued that within the use of communicative activities, there needs to be a 'Focus-on-Form'. They propose that, even though learners may be participating in interactions, with meaning as primary, they have some concern for

---

[1]Krashen and Terrell (1983) distinguished between 'learning' and 'acquisition'. They defined acquisition as an unconscious process that involves the naturalistic development of language proficiency through understanding and using language for meaningful communication. Learning, according to Krashen and Terrell (1983), was a process in which conscious rules about a language are

form. In effect, naturalness of communication is not compromised, but form has some priority. This move has been called "Focus on Form" (F on F) and is in contrast to traditional approaches of "Focus on Forms" (Long, 1991). The distinction between the two can be defined by referring to the fact that in F on F, instruction is not based on a linguistic form, but in order to satisfy learners' needs, it attempts to focus on a linguistic form. While in Focus on Forms, the primary purpose of teaching is linguistic forms as isolated from the context of language use. F on F, in effect, involves an occasional shift in attention to linguistic code features and is triggered by perceived problems with comprehension or production (Long & Robinson, 1998). As this principle has had a great influence on the current theories of language learning and teaching, emergence of task-based approaches to SLA research and language teaching, it should be defined and evaluated in this chapter. In order to have a broader perspective towards the context within which task-based instruction emerged, the relevant principles of "F on F" is discussed in the following section.

2.3 Focus on Form

In the previous sections, it was discussed how traditional approaches to language teaching were criticized and ultimately rejected. One tenable discussion was that teaching a language through its structure would not necessarily lead to the ability to use the language for communication. In fact, it was argued that what would push learners forward in language acquisition was the focus on meaning. However, proponents of "Focus on Form" argue that a mere focus on meaning would not lead to successful language acquisition. They propose that optimal language learning would

---

developed. Learning results in explicit knowledge about the forms of a language and the ability to verbalize this knowledge.

be achieved when not only is the language learnt through communicating with others but also when there is a *focus on form*, as opposed to *focus on forms*.

2.3.1 Rationale for Focus on Form

As clearly discussed earlier in this chapter, the communicative approach emerged in contrast to synthetic syllabi that focused on the presentation and practice of linguistic items, i.e., Focus on Forms (Long & Robinson, 1998). Communicative approaches accompanied by analytic syllabi, on the other hand, introduced a new trend in language teaching in which a focus on meaning substituted a Focus on Forms. Long and Robinson (1998) argue that this approach is equally single-minded since its focus on one aspect of language acquisition ignores the significance of the other. The inefficiency of such monolithic, i.e., meaning-focused, approaches to language teaching became more controversial when immersion programs proved to be ineffective in teaching a second language (Leeser, 2004; Swain, 1991). Although the Comprehensible Input Hypothesis hypothesized that meaningful input was the essential requirement for naturalistic acquisition of language, the majority of the students in immersion programs did not achieve adequate levels of accuracy in their use of morphology and syntax (Swain, 1991, 1993). It was argued that the learner's attention should be attracted by and directed to specific formal aspects of the language code in the context of meaningful use (Doughty & William, 1998; Long, 1999).

Criticizing approaches to language instruction that only focus on meaning, Long and Robinson (1998) point out a number of major problems resulting from such meaning-focused instruction. Firstly, they argue that exposure to comprehensible language samples alone is not sufficient for older language learners to learn the language successfully. Due to the physiological and psychological factors they are not able to

learn a language simply from exposure to it. Referring to persistent problematic language features, Long and Robinson (1998) further argue that there is evidence indicating that such exposure to language use would not enhance language acquisition. They also claim that formal language instruction can make L2 learning more efficient.

Norris and Ortega (2000) in a recent review of the research on L2 instruction have investigated the effectiveness of formal L2 instruction published between 1980 and 1998. The results of their studies indicate that focused L2 instruction results in large gains. These results have also revealed that explicit types of instruction are more effective than implicit types and the overall instruction appears to be more endurable.

2.3.2 Focus on Form vs. Focus on Forms

While discussions were being carried out about focus on forms and focus on meaning, Long (1991) distinguished between a *focus on forms* and a *focus on form*. He defined the former as an instruction that isolates or extracts linguistic features from context or from communicative activity. On the other hand, he defined the latter as not focusing on individual linguistic forms but as occasionally focusing on linguistic forms triggered by an analysis of the learners' needs. He maintained that focus on form " … overtly draws students' attention to linguistic elements as they arise incidentally in lesson whose overriding focus is on meaning or communication" (Long, 1991, p. 45-46). Extending the concept of focus on form, Long and Robinson (1998) contend that

" focus on form involves … an occasional shift in attention to linguistic code features – by the teacher and/or one or more students – triggered by perceived problems with comprehension or production" (Long & Robinson, 1998, p. 23). However, it is pointed out that any shift of attention from meaning processing to form processing

should be a brief response to problems in 'on-line communication' (Doughty, 2001). In other words, pedagogical interventions should not interrupt language use or interfere with the larger macroprocessing involved in speech comprehension or production.

Doughty and William (1998) further distinguish, but not contrast, the concepts of focus on forms and focus on form. They argue that:

> Focus on forms and focus on form are not polar opposites in the way that 'form' and 'meaning' have often been considered to be. Rather, a focus on form entails a focus on formal elements of language, whereas focus on forms is limited to such a focus, and focus on meaning excludes it. Most important, it should be kept in mind that the fundamental assumption of focus-on-form instruction is that meaning and use must already be evident to the learner at the time that attention is drawn to the linguistic apparatus needed to get the meaning across. (Doughty & William, 1998, p. 4)

Although these definitions may vary, to some extent, from one another, they all refer to the same underlying principles of F on F. As Ortega (1999) summarizes, all these definitions presuppose (a) that the learner be engaged in meaning before attention to the codes occurs; (b) that the choice of form(s) to be targeted take into account developmental considerations of learner readiness; and (c) that the targeted form(s) be an integral part of the communicative needs engendered by the instructional language use event.

Skehan (2003) argues that there is an agreement about F on F among different approaches to task-based instruction. However, they differ in how they try to achieve such a focus on form: some rely on feedback (Doughty & Williams, 1998), some on attention allocations (Skehan 1998), some on interaction (Van Lier & Matsu, 2000)

and some on output (Swain & Lapkin, 2001). Hence, they may approach F on F at different stages of teaching or through different processes.

2.3.3 Focus on Form and Language Teaching

In order for the language to be acquired and used, as it goes almost without saying, language must be cognitively processed. Cognitive constructs and processes like memory, speech processing, encoding and decoding are indispensably involved in the whole process of comprehending and producing language (Doughty, 2001; Levelt, 1989). It is argued that F on F procedures can potentially influence these processes, but only if the intervention conforms sufficiently to the nature of language encoding underway (Doughty, 2001). It is believed that F on F will facilitate language learning processes if the intervention manages not to disrupt or halt the fundamental and ongoing cognitive macroprocessing that comprises comprehension and production (Doughty, 2001; Doughty & William, 1998). This means that if a learner is disrupted for correction purposes while encoding a message for production, the process of encoding would break down at the point of intervention.

There is a lot of discussion about when and how to implement F on F. Long (1991) considers F on F as a reactive process in which the teacher draws attention to particularly salient errors made by learners while communicating a message. Proposing a more proactive approach to F on F, Doughty and William (1998) argue that the reactive approach is too demanding for the teacher and makes teaching difficult. This requires the teacher, based on experiences and observations of a particular group of learners, to predict or determine what learning problems to focus on.

There is an agreement, in second language acquisition research, that F on F should ideally come at cognitively favorable times when the intervention can be seamless with processing for language learning (Doughty & William, 1998; Long & Robinson, 1998). Doughty (1997) cites three criteria to be met if a pedagogical intervention is to be considered unobtrusive. The three criteria are (a) the primary focus should be on meaning; (b) the focus on form targets should arise incidentally; and (c) learner attention should be drawn to form briefly. Further discussions of when and how F on F is considered as favorable by different approaches to TBI will be provided later in this chapter.

As indicated in the discussions presented in the previous sections, F on F enjoys great significance in second language acquisition research and in language pedagogy. As regards this study, F on F is directly related to and reflected in the underlying principles of TBI, which will be discussed in the following sections.

2.4 Emergence and Principles of Task-Based Instruction

During the 1980s and 1990s advocates of communicative approaches to language teaching (Brumfit & Johnson, 1979) and SLA researchers (Long & Crookes, 1992) started criticizing the current approaches to the language teaching of the time. Each of these groups argued that interlanguage development will come about, not through control and practice of language items, but through the meaningful use of language and the engagement of more naturalistic acquisitional processes. Second language teachers and researchers, frustrated by the limitations of the previous approaches to language teaching, showed an enormous interest in task-based language instruction during the 1980s and 1990s. This shared interest was motivated to a considerable extent by the fact that "task" was seen equally important to SLA researchers and to

language teachers (Bygate, 1996; Candlin, 1987; Pica, 1997; Skehan, 1996a, 1998; Swain, 1985; Robinson, 1995; J. Willis, 1996). Being inspired by the innovative principles of the communicative approach in general and by communicative activities, i.e. tasks, in particular, researchers and teachers began experimenting with task-based instruction in the early eighties. The success of Prabhu's Communicational Teaching Project (Prabhu, 1987) in Banglore encouraged more researchers and pedagogues to follow task-based language teaching. As described by Prabhu (1987), his project attempted to implement an approach to language teaching which focused on the ability to perform a task or activity, and avoided the explicit teaching of grammatical structure. It is worth mentioning that, despite being an innovative and praiseworthy program employing tasks as units of language teaching, Prabhu's project has been criticized because it focused more on task completion rather than on the language used in the process. In addition, the main complaint about Prabhu's project is its failure to build an evaluation component into the design of a task-based approach to TBI (See Long & Crookes,1992, for a detailed discussion).

The notion of task, as Leung et al. (2004) argue, has a long tradition. However, more recently, task has been used to refer to meaning-oriented learning activities which require interaction with others. A primary definition of task was provided by Nunan (1989) as " a piece of classroom work which involves pupils in comprehending, manipulating, producing or interacting the target language while their attention is principally focused on meaning rather than on form" (p. 10). Detailed definitions and discussions of task will be presented later in this chapter.

Task-based instruction, as a general term, refers to different types of teaching situations in which tasks are employed, as means of instruction, to activate, stretch and refine current interlanguage resources and processing capacities (Samuda, 2001).

The assumption in employing tasks is that "transacting tasks in this way will engage naturalistic acquisitional mechanisms, cause the underlying interlanguage system to be stretched, and drive development forward" (Skehan, 1998, p. 95). In TBI, tasks are communicative activities that drive the learner's interlanguage system forward by engaging acquisitional processes (Long & Crookes, 1992). As Samuda (2001) argues, although the degree of prominence tasks occupy in an overall programme of language instruction may vary, the standard pedagogic purpose mainly associated with using tasks in TBI is the provision of communication practice, through which language-processing capacities may be developed in the context of language use.

TBI has a number of purposes. J. Willis (1996, pp.35-6) identifies eight purposes which mainly relate to communicative effectiveness and L2 acquisition:

1. to give learners confidence in trying out whatever language they know;
2. to give learners experience of spontaneous interaction;
3. to give learners the chance to benefit from noticing how others express similar meanings;
4. to give learners chances for negotiating turns to speak;
5. to engage learners in using language purposefully and co-operatively;
6. to make learners participate in a complete interaction, not just to one-off sentences;
7. to give learners opportunities to try out communication strategies; and
8. to develop learners' confidence that they can achieve communicative goals.

The TBI approach views learning theory from a rather different perspective. It emphasizes the fact that language input, however provided, simply offers raw materials on the basis of which learners may review their picture of the target language system. Learning, in TBI, is seen as a developmental process through which

a learner moves towards the target form by engaging in meaning and through learning by doing. TBI is believed to be essentially directed at both improving students' abilities to use the target language and enabling them to acquire new linguistic skills. Errors are considered inevitable and are a common occurrence in learning regardless of the learners' background language and knowledge. As indicated earlier in this section, a distinctive feature of TBI is that the focus on the language form comes at the end. D. Willis (1996) distinguishes a task-based approach from presentation methodologies (widely known as the PPP method representing presentation, practice and performance) by the priority they give to accuracy as opposed to fluency, or to the form rather than to the meaning. In effect, in a task-based approach learners begin with a holistic experience of language in use and only later may they have an opportunity of having a closer look at some of the grammatical features naturally occurring in that language. Robinson (2001) argues that task-based approaches differ from other approaches to language teaching because they have a performance emphasis and are not predicated on the assumption that levels of target-like accuracy will be achieved as a result of practising specific structures. He further contends that achievement in TBI is accounted for by the results obtained in terms of performance, and not regarding the language system. The last difference, Robinson (2001) points out, is that in TBI exposure to language form predominantly takes place in the context of communicative activities; while in other types of instruction, exposure to form occurs in isolation.

To summarize, many language teachers and researchers in TBI appear to be moving away from the traditional focus on forms. They are seeking an approach that both promotes the learner's ability to interact to achieve communicative goals in the real

world and facilitates L2 acquisition by drawing the learner's attention to linguistic forms and developing language processing capacities.

There are different research oriented approaches to TBI, each with certain principles and theoretical assumptions. In the section that follows, I will introduce some of these approaches and will in detail discuss the underlying principles and theoretical assumptions of the approach that is adopted in this research.

2.5 Research Oriented Approaches to Task-Based Instruction

A number of different approaches to task-based instruction have been distinguished within the current SLA literature. In this section, three influential approaches to TBI are introduced and examined. These approaches are different theoretical accounts of task-based language instruction which are currently employed by many researchers and pedagogues. The three approaches are: 1) a psycholinguistic approach; 2) a socio-cultural approach and 3) a cognitive approach.

2.5.1 A Psycholinguistic Approach

This approach, which has been essentially influenced by Long's (1983, 1989) Interaction Hypothesis, represents the first major moves that have emerged in TBI. In its early form (Long, 1983, 1989), the Interaction Hypothesis claimed that acquisition is facilitated when learners obtain comprehensible input as a result of the opportunity to negotiate meaning when communication breakdown occurs. In its later form (Long, 1996), the theory is extended to explain other ways in which negotiation of meaning can contribute to L2 acquisition. The Interaction Hypothesis proposes that the negotiation of meaning, in fact, may occur through the feedback that the learners receive on their own production when they attempt to communicate, or through the

learners' modified output and reformulated production. In this way, negotiation of meaning serves to draw learners' attention to linguistic form in the context of a primary focus-on-meaning and 'noticing' that is perceived to be necessary for acquisition to take place (Schmidt, 1990).

Negotiation of meaning concerns the way learners encounter difficulties that arise in their communication with others while they are engaged in tasks. Long (1996) proposes that the interactional adjustments that learners employ to address such difficulties would encourage their interlocutors to modify the input they are providing. More significantly, the responses learners receive while negotiating meaning would deliver feedback to the learner at a highly favorable moment. As Pica (1994) points out, this feedback arises when meaning is problematic and when the learner is thought to be most receptive. Long (1989) contends that tasks which generate beneficial negotiation of meaning of this sort are indexed by greater number of clarification requests, comprehension checks, and confirmation checks. These conversational features are taken to indicate the degree of usefulness of the interactions involved in performing the task. In more recent accounts of the Interaction Hypothesis, a new conversational feature is being introduced, i.e. recast. Recast, in this sense, refers to an instance when an interlocutor phrases something said by a learner, and so provides a model and feedback when s/he may be most open to such a contribution (Long, Inagki & Ortega, 1998). To put it in another way, recast is repetition of a learner's incorrect utterance, but with changes made to make it correct.

A number of effects of task characteristics and task conditions on negotiation of meaning and recasts have been revealed through this approach to tasks. Research, in this domain, suggests that convergent tasks – i.e. tasks in which participants have to agree on an answer- would produce more negotiation of meaning than divergent tasks

– i.e. tasks where no agreement is necessary to be arrived at (Long, 1989). A study by Duff (1986) investigated this possibility and provided partial supportive results. Pica and Doughty (1985) showed that group and pair-fronted interactions provided more modifications than a teacher-fronted interaction. Mackey, Gass, and McDonough (2000) and Lyster and Ranta (1997) have reported significant recasting in both classroom-based and experimental studies. Nicholas, Lightbown and Spada (2001) propose that recasting is more effective when a learner has already begun to use a language feature so that s/he is able to discriminate between alternatives.

The negotiation of meaning and recasts studies have made a great contribution to TBI by providing a wider perspective of task characteristics and conditions. However, this approach to TBI has received some criticism, as well. Aston (1986) reported that tasks that require a lot of negotiation of meaning are not favorable to learners. Foster (1998) argues that in the typical classroom context negotiation of meaning does not occur frequently. She adds that when such negotiations of meaning happen they are primarily lexical and are not accounted for by particular students. Lyster (1998) criticizes the emphasis the Interaction Hypothesis places on recast. He argues that learners do not usually notice recast, and even when they do it is not incorporated in the learner's speech. Finally, Ellis (2000) criticizes the negotiation of meaning research because it provides little information about how different task characteristics interact with the impact they have on negotiation of meanings.

2.5.2 A Socio-Cultural Approach

This approach to task-based instruction draws on socio-cultural theory which has grown out of the work of Vygotsky (1978) and his followers. Socio-cultural theory is another approach to interaction in language teaching situations where teachers and

researchers attempt to explore how learners co-construct meaning while they are interacting with each other. In effect, one central claim of socio-cultural theory is that learners always co-construct the activity they are engaged in, in accordance with their own socio-history and locally determined goals (Lantolf, 2000). As Appel and Lantolf (1994) discuss performance depends crucially on the interaction of individual and task rather than on the inherent properties of the task itself. It is stated in socio-cultural theory that the same task can result in very different kinds of activity when performed by the same learners at different times. Coughlan and Duff (1994) report that the same task has been performed differently by the same participants on two different occasions. They show how an entire range of discourse types arose from this task reflecting their subjects' multiple interpretations of it. They conclude that 'tasks' can not be treated as a constant in research as "the activity it generates will be unique" (p.191). In fact, it is assumed that the interest in tasks is to allow participants to shape it to their own ends and to build meanings collaboratively.

Learning, according to socio-cultural theory, arises not *through* interaction but *in* interaction. Learners first succeed in performing a new function with the help of another person and then internalize this function so that they can perform it independently. In this way, social interaction mediates learning. The kinds of interactions that most successfully mediate learning are those in which the new functions are 'scaffolded' by the participants. 'Scaffolding', in this sense, refers to the dialogic processes by which one speaker assists another to perform a new function. As Wood, Bruner and Ross (1976) discuss, scaffolding can involve recruiting interest in the task, simplifying the task as necessary, maintaining pursuit of the goal of the task, or controlling frustration during problem solving.

Task-based research in the socio-cultural perspective has been directed to show how scaffolding might help learners achieve a successful task outcome. The following are examples of the work conducted within this perspective. Donato (1994) describes how the collective scaffolding employed by a group of learners enabled them to produce a grammatical structure jointly even though none of the students knew it individually. Samuda (2001) shows how a teacher created the conditions for students to uptake a new grammatical feature through implicit scaffolding. Van Lier and Matsu (2000) were interested in the nature of interaction within the socio-cultural theory. Through their study, they explored whether interactions vary measurably in how symmetrical and collaborative they are. They examined whether interlocutors ratified one another's contributions, whether they responded to these contributions and developed them, or, in contrast, whether they failed to notice these things when one interlocutor dominated the whole interaction. In their research, Van Lier and Matsu (2000), found clear differences between learners on these indices.

To summarize, the socio-cultural approach to TBI has focused on how tasks are accomplished by learners and teachers and how the process of accomplishing them might contribute to language acquisition. Proponents of the socio-cultural approach to TBI view the learner, the teacher and the setting in which they interact as significant as the task itself. They mainly focus on how participants achieve inter-subjectivity with regard to goals and procedures. They are keen to know how the participants collaborate to scaffold each other's attempt to perform a task which lies outside their individual abilities.

The socio-cultural approach to TBI has been criticized for some of its assumptions. The main criticism is that it has concentrated on describing the social interactions that arise when learners perform tasks and have made little attempt to indicate whether

these interactions contribute to L2 acquisition (Ellis, 2000). In rejecting the deterministic view of tasks made by the psycholinguistic approach to tasks, socio-cultural researchers have failed to acknowledge that task characteristics and conditions do impact on task performance, although they may not be precisely specified. As socio-cultural theory anticipates a minimized systematic result emerging from the learners' performance on a particular task at different times, it will be difficult to make reliable predictions regarding the kinds of language use and opportunities for learning that will arise from a particular task. I will not attempt to further investigate the principal concepts and different aspects of socio-cultural theory since they are not germane to the research design of this study.

2.5.3 A Cognitive Approach

The third approach to TBI to be introduced and examined in this section is a cognitive approach. Essentially, this approach to TBI draws upon principles of the current theories of cognitive psychology. This approach predominantly argues that as language acquisition occurs through cognitive processes such as speech and information processing, encoding and decoding, the cognitive difficulty of a task has significant implications for understanding how attention is deployed through task completion. Knowing what demands the task will make opens up the possibility of using task design to manipulate the learner's attention in ways that may help interlanguage development.

The cognitive approach to TBI is based on a distinction in the way in which learners are believed to represent L2 knowledge. Second language learners, while acquiring a language, construct an exemplar-based system as well as a rule-based system. The former is lexical in nature and includes both discrete lexical items and ready-made

formulaic chunks of language. The linguistic knowledge in this system can be easily accessed and therefore ideally suits fluent language performance. The rule-based system consists of the abstract representations of the underlying patterns of the language. They require more processing and thus are best suited for more controlled language performance. The rule-based system is needed when learners have to creatively construct utterances to express meaning precisely in socio-linguistically appropriate ways. The distinction between these two types of linguistic knowledge is clearly acknowledged in cognitive psychology (N. Ellis, 1996) as well.

According to this perspective, tasks are viewed as devices that provide learners with the data they need for learning. In a task, meaning is primary and the activity is outcome-evaluated (Skehan, 1996). The authenticity of a task would further help learners develop their interlanguage system and be prepared for the real-world use of language. The design of a task is seen as potentially determining the nature of the language use and opportunities for learning that arise. Hence, teachers' or testers' choice of a task will not be a neutral matter (Skehan and Foster, 1997).

Researchers (Mehrnet, 1998; Foster and Skehan, 1996; Robinson, 1995; Skehan, 1998) who have taken a cognitive perspective towards tasks have focused on the psychological processes that are typically involved when learners perform tasks. Researchers in this field have explored three main areas: 1) analyses of how attentional resources are used during task completion; 2)the influence of task characteristics, conditions and design on language performance; and 3) the impact of task selection and use on learning.

Researchers from a cognitive approach to TBI have focused on the learners' production in order to find what task characteristics would influence task difficulty. Learners' production has been investigated in terms of accuracy, complexity, fluency

and lexical variety. Results of a number of studies have shown that, with tasks of different designs and characteristics, learners produce language of different quality. Foster and Skehan (1996, 1999) have reported that interactive tasks lead to more accuracy and complexity while monologic tasks lead to more fluency. Crookes (1989) and Ellis (1987) have shown how pre-task planning influences fluency and accuracy of language performance. Bygate (1996) has reported the positive influence of task repetition on different aspects of performance. A detailed account of the variables investigated in task-based studies will be provided in Chapter IV.

There are two contrasting approaches within the cognitive approach to TBI, but they also share many similarities. Skehan (1998) proposes that attentional resources are limited, and that to attend to one aspect of performance (complexity, accuracy, or fluency) may well mean that other dimensions suffer. Skehan and Foster (1997, 2001) argue for the existence of tradeoffs in performance, such that, typically, greater fluency may be accompanied by greater accuracy or greater complexity, but not both. The second approach is presented by Robinson (2001) who advocates two propositions: (a) that attentional resources are not limited in the way Skehan (1998, 2001) argues, but instead learners can access multiple and non-competing attentional pools, and (b) that complexity and accuracy in a task correlate, since they are each driven by the nature of functional linguistic demands of the task itself. Details of these contrasting views will be discussed in a later section in this chapter and the relevant discussions will be further developed in Chapter IV.

From a cognitive and neurolinguistic perspective, cognitive processes seem to form a major part of second language acquisition (Levelt, 1989). It is certain that cognitive activities and processes like memory and speech processing are an important part in the whole process of comprehending and producing language. From a TBI

perspective, the strength of the cognitive approach is that it has served to identify features of task design, task characteristics and performance conditions which can impact on L2 performance and L2 acquisition. However, there are two main criticisms to the cognitive approach to TBI. One drawback is that, to date, there has been no single general measure of task performance to be used to determine task difficulty (Ellis, 2000). This implicitly indicates that more research is required to enable researchers to provide clear and general measures of task difficulty. The second criticism, which shares in part a similar standpoint of the socio-cultural theory to the research carried out within the cognitive approach concerns the lack of attention to the learners' variables, i.e. learners' perceptions and attitudes to tasks and task difficulty. Although some researchers (Robinson, 2000) have begun to consider this point, learners' perceptions of task difficulty have mainly remained unexplored.

My general interest in task-based pedagogy and assessment embraces all different approaches and perspectives currently represented in the relevant research literature (see above). However, in this research study, I will be paying particular attention to the cognitive approach because I would like to investigate some of the key task characteristics and properties in a way that would allow statistical measurements. Moreover, as I will be investigating tasks in a language testing context, the cognitive information processing perspective would offer a more reliable and promising framework for considering and estimating task difficulty and its effect on language performance. It is hoped that the outcomes of this study will contribute to a greater understanding of TBI and TBA from a quantitative point of view.

2.6 Task

2.6.1 Definition

As Bygate et al. (2001) point out, definitions of tasks are generally diverse and therefore not broadly agreed upon. Researchers of different theoretical perspectives to TBI have defined "task" in a number of different ways. Long (1985) defines task as " a piece of work undertaken for oneself or for others, freely or for some reward. Thus examples of tasks include painting a fence, dressing a child, filling out a form, … . In other words, by 'task' is meant the hundred and one things people do in everyday life, at work, at play, and in between" (p. 89). J. Willis (1996) describes task as " a goal-oriented activity in which learners use language to achieve a real outcome" (p. 53). Learners, in Willis' opinion, may use whatever target language resource they have to solve a problem, do a puzzle, play a game, or share experiences. She further clarifies that language activities that focus on rehearsal of linguistic forms and do not have a real-life outcome are not considered as tasks.

Nunan (1989) reviews various definitions of task and reports that they all agree that "tasks involve communicative language use in which the user's attention is focused on meaning rather than linguistic form" (p. 10). He defines task as "a piece of classroom work which involves learners in comprehending, manipulating, producing or interacting in the target language while their attention is principally focused on meaning rather than form" (p. 10). He argues that tasks should represent complete communicative acts and be analyzed or categorized in terms of their goals, input data, activities, settings and roles.

Ellis (2000, 2003) views a task as a 'workplan'. That is, it takes the form of the materials for researching or teaching language. As he puts it, a workplan typically involves (a) some input, i.e. information that learners are required to process and use,

and (b) some instructions relating to what outcome the learners are supposed to achieve.

Skehan (1996 a, 1998), following Candlin (1987) and Long (1989), proposes that task is "an activity in which meaning is primary; there is some communication problem to solve; there is some sort of relationship to comparable real-world activities; task completion has some priority; and the assessment of the task is in terms of outcome" (p. 38). This definition has reflected a broad consensus among and has been followed by many researchers and educators (Foster and Skehan, 1996; Mehnert, 1998; Ellis, 2000). The discussions of 'task' clearly indicate that tasks are distinguishable in terms of characteristics and performance conditions. For instance, they may involve different language skills; they may draw on learners' input, knowledge or experience; they may be based on written or spoken texts; they may be drawn from visual data; they may be performed by one, two or a group of learners; or a number of other different characteristics. However, the most significant point central to performing tasks is that while doing the tasks learners will focus on meaning. They use the language to exchange meaning for a real purpose and they are free to use whatever language structures they want. A fundamental principle of performing tasks, in TBI, is that the use of language would replicate features of real language use outside the classroom (Long, 1989; Prabhu, 1979). Another salient feature of TBI is that tasks are eventually outcome-evaluated. In effect, task performance is evaluated through investigating whether the learners or participants have achieved the outcome of the task.

2.6.2 Task Difficulty and Task Sequencing

The early discussions of task difficulty, task selection and sequencing are originated from research being carried out in syllabus design. It was primarily assumed that in a task-based approach to language teaching, tasks, rather than grammatical structures or linguistic criteria are to be used as basic units of language teaching syllabi. The general concern of syllabus design, therefore, should be determining the order with which tasks should occur in a syllabus. Hence, it is crucial to be able to assess task difficulty in order to select and sequence tasks for teaching as well as assessment purposes (Norris et al., 2002).

Skehan (1998), adopting a cognitive and attention-driven perspective, contends that the purpose of determining task difficulty is twofold. The first reason is that tasks of appropriate difficulty are likely to be more motivating to learners. The second reason, as he states, deals with the attentional capacities of human mind. As such capacities are limited, he argues for the use of tasks of appropriate difficulty so that learners will be able to cope with the demands upon their attentional resources. If a task of appropriate level of difficulty is selected, there will be much greater likelihood that noticing will occur, that balanced language performance will result, and that spare attentional capacities can be challenged effectively (Skehan, 1998).

Long (1991, 1996) places great importance on carrying out a needs analysis to obtain an inventory of target or real-world tasks that a particular group of learners will undertake. In line with the psycholinguistic approaches to interaction, Long (1996) considers "negotiation of meaning" another necessary criterion for the selection of the tasks. As explained before, negotiation of meaning concerns the way learners encounter communication difficulties while completing tasks, and subsequently employ some means to deal with the difficulties. In fact, Long (1989) proposes that

the amount of "negotiation of meaning" a task would generate when learners are communicating during the task has a determining role in selecting and sequencing tasks in a task-based syllabus. That means, tasks with more negotiation of meaning - i.e. more clarification requests, clarification checks and confirmation checks – are taken to be more supportive of acquisition. As a result, two-way tasks, i.e. tasks that require a two-way exchange of information, produce more transfer of meaning and are more opportune for the learning purposes.

Along with the choice of tasks on the basis of their difficulty there is a discussion on how to sequence tasks in a syllabus. As Robinson (2001) puts it, a syllabus can consist of a prospective decision about what and in which order to teach. In such a case, the syllabus will be a list of the classroom activities. Another type of sequencing may be implemented in terms of the on-line decisions about the content (Breen's process syllabus, 1984), in which case, the initial syllabus will only guide, but not restrict, the classroom activities. The last type of sequencing, as Candlin (1984) proposes, would be retrospective sequencing in which no syllabus will emerge until after the course of instruction. In this type of sequencing, a syllabus functions only as a record of what has been done.

Broadly, in task-based approaches, sequencing is mainly based on a prospective decision about the increasing difficulty of tasks for the learner. As Skehan and Foster (2001) contend "knowing what demands the task will make opens up the possibility of using task design to manipulate the learner's attention between form and meaning in ways that may help IL development" (p. 194). They argue that tasks can be best categorized to reflect their cognitive load. Once this is done, they add, these categories might be used in planning a task-based approach to language teaching.

However, in order to determine the difficulty level of a task, various characteristics of tasks are to be investigated and an index of task difficulty to be explored. In the next section different task characteristics and their effect on the difficulty of a task will be investigated and discussed.

2.6.3 Task Characteristics

Since the early days of the emergence of TBI, SLA researchers have been investigating different characteristics of tasks to determine various aspects of task difficulty. A number of proposals have been made in this regard. A brief overview of the relevant research on task characteristics and task difficulty will be presented in this section.

Brindley (1987) proposes one of the earliest classifications of the factors affecting task difficulty. He distinguishes learner, task, and text factors as the three significant elements in task difficulty. Brindley's text factor is not relevant to the focus of this research. His learner factors include confidence and motivation, along with prior learning experience, ability to learn at the pace required, and possession of necessary language skills and relevant cultural knowledge. Brindley (1987) argues that the presence of all these elements in a task would make performing the task easier to the learner. This claim directly touches upon the notions of task difficulty. However, it doesn't seem to be useful in determining task difficulty and in making decisions about sequencing of the tasks. As one of the primary investigations of task difficulty, Brindley's scheme has contributed to our understanding of task difficulty. But there are other characteristics and processes inherent to tasks and TBI which are not accounted for in Brindley's scheme.

Another significant early contribution to the issue of task characteristics and task difficulty resulted from Prabhu's work (1987). Prabhu, working on the Banglore Communicative Project, attempted to employ theoretical rationales and practical procedures of the communicative approach to develop a framework for the selection and sequencing of the tasks. Through employing a task-based 'procedural' syllabus, rejecting the linguistically graded materials, and avoiding explicit explanation of the rules or corrective feedback, Prabhu approached the problem of assessing the difficulty of a task. As a result of observing which tasks were most effectively used and more successfully generated communication among learners, he recommended reasoning-gap tasks above all, in preference to opinion-gap and information-gap tasks. An example of such a task is the learners' planning a railway journey across India, armed with railway timetables and schedules.

The tenets of Prabhu's work were later criticized because of ignoring the importance of noticing in language learning processes, as mentioned by Schmidt (2001), and lack of a F on F (Long, 1991; Doughty, 2001). Another area, which was not addressed in Prabhu's scheme, was the interplay between the cognitive demands of a task and the task conditions. In other words, Prabhu's scheme did not consider whether changing the nature of the solution to the task would interact with the cognitive demands of a task.

Another early and highly influential attempt to characterize task difficulty was made by Candlin (1987). He proposed a set of criteria for selecting and grading tasks. The first characteristic he mentioned was the cognitive load of a task. By this he referred to the general complexity of the content of the task including the number of participants or elements of the task or the naturalness of the sequence it may require to follow. Communicative stress was the second feature, which referred to the amount

of stress a task may have for the participant with regard to the interlocutor, their language proficiency or knowledge. Code complexity was another characteristic of a task and represented the complexity of the linguistic items of a task, which was assumed to be directly influencing task difficulty. In direct connection with code complexity was the interpretative density that represented the linguistic and argumentative complexity of the texts used on tasks. In addition, particularity and generalizability were two contrasting aspects of task characteristic in Candlin's framework. These two terms referred to clarity, novelty and specificity of the goal of the task and the norms of interpretation. The last task characteristic in Candlin's framework for task difficulty was process continuity which was derived from the familiarity of the task type and the capacity of the learners to cope with unfamiliar tasks.

Candlin's definitions of task characteristics and task difficulty received support from different theories of cognitive psychology and information processing. His framework seemed to be the most comprehensive framework for task characteristics and task difficulty of the time. However, as criticized by a number of researchers, it offered no transparent guidelines to materials and syllabus designers and was non-complementary in many ways (Skehan, 1998; Robinson 2001).

Skehan (1996, 1998) has attempted to overcome this shortcoming and has proposed a similar framework for analyzing and categorizing task characteristics and task difficulty. Skehan has proposed a three-way distinction for the analysis of tasks, based on code complexity, cognitive complexity and communicative stress. He contends that his categorization "groups some of the factors suggested by Candlin into slightly higher-order categories" (Skehan, 1998, p. 99). The full scheme he proposes is as follows:

a) Code Complexity

- linguistic complexity and variety
- vocabulary load and variety
- redundancy and density

b) Cognitive Complexity

Cognitive Familiarity

- familiarity of topic and its predictability
- familiarity of discourse genre
- familiarity of task

Cognitive Processing

- information organization
- amount of computation
- clarity of information given
- sufficiency of information given

c) Communicative Stress

- time limits and time pressure
- speed of presentation
- number of participants
- length of text used
- type of response
- opportunities to control interaction.

It should be explained that in Skehan's scheme the major constituents are the *language requirement*, the *thinking requirement* and the *performance-condition requirements* of a task. In fact, he has distinguished these three factors from one another that is different from Candlin's proposal. An interesting point in this scheme

is the distinction made between cognitive familiarity and cognitive processing requirements of the tasks. In his work, Skehan (1998) defines cognitive familiarity as "the capacity to access 'packaged' solutions", and contrasts it with cognitive processing, "the need to work out solutions to novel problems 'on-line'" (p. 99). This reflects the fact that familiarity with the topic, the genre or the task might facilitate the process of performing the task because a familiar task requires existing chunks of knowledge to be retrieved and employed in performance. On the other hand, there is another cognitive aspect, processing, which would require the learner to achieve some solutions while performing the task. In this case, the attentional resources are stretched since the processing has to be directed at the cognitive problem involved in the task.

As discussed before, Skehan's framework for task difficulty is built on the principles of a cognitive approach to language learning. He (2001) argues that "humans have limited information processing capacity and must therefore prioritize where they allocate their attention" (p. 189). In other words, if a task needs a lot of attention to its content, for example it is complex or puzzling, there will be less attention paid to its language.

Within a similar cognitive framework, Robinson (2000, 2001) proposes a triadic framework for investigating task difficulty, which would consequently determine the selection and sequencing of tasks. He distinguishes three interacting groups of factors that influence task performance and learning: the cognitively defined task complexity, learner perceptions of task difficulty, and the interactive conditions under which tasks are performed. The details of Robinson's scheme are as follows:

a) Task Complexity (Cognitive factors)

- resource-directing e.g. +/- few elements; +/- here-and-now; +/- no reasoning demands

- resource-depleting e.g. +/- planning; +/- single task; +/- prior knowledge

b) Task Conditions (Interactive factors)

- participation variables e.g. one-way/two-way; convergent/divergent; open/closed

- participant variables e.g. gender; familiarity; power; solidarity

c) Task Difficulty (Learner factors)

- affective variables e.g. motivation; anxiety; confidence

- ability variables e.g. aptitude; proficiency; intelligence

Task complexity, as Robinson (2000, 2001) argues, is the result of the attention, memory, reasoning, and other information processing demands imposed by the structure of the task on language learner. His task complexity mainly consists of two dimensions manipulated by task design: resource-directing and resource-depleting dimensions. Resource-directing dimensions are those aspects of a task which require more reasoning or more information transition. Increasing task complexity along a resource-directing dimension, in Robinson's framework, would make tasks more demanding which can be met by using specific features of language code. In contrast, a resource-depleting dimension makes greater demands on attention and working memory, but does not direct the resources to features of language code. This instance happens when a secondary task is added to the first or when learners have to do a task without pre-task planning time. Robinson's task complexity involves a number of previously identified task factors suggested by Skehan's (1996, 1998) cognitive complexity.

Task difficulty, in Robinson's framework, refers to a series of learner factors which may make a task more or less difficult. These factors are, in fact, differentials in the resources available to the individual learners including attentional, memory, and reasoning resource pools. Task difficulty, in this context, will define the-between-learners differences and includes affective variables (confidence, motivation, anxiety) as well as ability variables (intelligence, aptitude, cognitive style).

Robinson's task condition involves neither task factors nor learner factors alone, but participation factors such as the direction of information flow and the communicative goals of task performance. Familiarity with other participants and with task role, task goal and task interpretation are all examples of task condition.

The framework presented by Robinson shares many of the cognitive features of Skehan's (1998) cognitive scheme of task characteristics. These two frameworks, in fact, attempt to indicate how attentional resources are used during task completion; how task characteristics influence learner performance; and how different conditions under which tasks are completed may influence performance.

Although they are representing a similar cognitive approach to task characteristics, they propose some contrasting issues. Skehan (1998) proposes that attentional resources are limited, and that to attend to one aspect of performance- complexity, accuracy or fluency of language – would hurt the other aspects. Studies carried out by Skehan and Foster (1997, 2001) support such a proposal by demonstrating the existence of tradeoffs in performance, i.e. greater fluency may be accompanied by greater accuracy or complexity but not both.

Robinson has, in contrast, proposed that attentional resources are not restricted in the way Skehan and Foster (2001) have argued. He believes learners can access multiple

and non-competing attentional resources. He further proposes that complexity and accuracy correlate with one another while they contrast with fluency.

Robinson's framework is also different from Skehan's framework with regard to the emphasis he puts on the learner variables. He is considering learners factors, e.g. motivation, anxiety, aptitude, as elements influencing task difficulty. However, it should be taken into consideration that learner variables exist in all language learning processes and interact with all various sub-processes. Although learner factors do play a significant role in any learning situation, owing to the complicated nature of human being as different individuals, they cannot be simply explored or measured. As a result of having numerous learner variables, individual differences arise and one learner seems to be more successful than another is. But whether these individual factors and differences could determine the concept of task difficulty and thus be taken in to account as principles of selecting and sequencing tasks needs more systematic research.

While both contrasting perspectives of Skehan and Robinson on issues of attentional resources and their impact on different aspects of performance seem viable, more research is inevitably required to investigate the very intricate nature of the relationship between the attentional resources and aspects of language performance. In fact, in the following chapters it will be explained that one purpose of this study is to investigate whether attentional resources are limited (as Skehan and Foster, 1997 and 2001 propose) or whether learners have access to multiple non-competing attentional pools (as Robinson, 2001, puts it).

Bachman (2002), reviewing Skehan's definitions of task difficulty, views code complexity a real feature of task difficulty but argues that cognitive complexity and communicative stress are functions of the interaction between the test-taker and the

task (p. 466). He believes that cognitive complexity could be seen not as different factors that affect language performance directly, but as interactions among other determinants of test performance, i.e. level of language ability, risk taking, cognitive style and affect. Undoubtedly, more research studies are needed to investigate whether, as Bachman argues, cognitive complexity is a function of the interaction between different determinants of test performance or a real feature of task difficulty.

In the chapter that follows, I will attempt to show how language testing has employed tasks and task-based assessment in developing appropriate tests that can evaluate learners' communicative language ability. In order to explain task-based language assessment, I will discuss different phases language testing has gone through, the key properties of a language test, models of language ability and examples of oral language tests.

CHAPTER III

Task-Based Language Assessment

3.1 Introduction

In the current context of language assessment, tasks are often deployed in language test development in order to elicit test-takers' language performance and to be used as the basis of evaluating test-takers' language ability. A large number of local and international organizations are using tasks for the evaluation of language ability including oral language ability (e.g. IELTS, University of Cambridge Examinations, TOEFL). Hence, the prime purpose of this chapter is to investigate how the use and evaluation of tasks in some international language tests are construed and justified from a language testing perspective. To achieve this purpose, I will first start with a short background to second language testing and its relationship to SLA. Then I will follow with a brief consideration of the three significant phases of language testing, as a way of differentiating task-based assessment from other types of language testing. Key features of language tests including validity, reliability, and authenticity of the tests and/or test results will then be discussed. The next section of the discussion will highlight the prominent models of language ability. As the purpose of the present research is directly related to the assessment of oral language ability, the issues relevant to the construct of oral language ability will be discussed and some common oral language tests will be examined. Task-based assessment will then be introduced

and evaluated. The chapter will conclude with important discussions about the key properties and the problematic features of task-based assessment (TBA).

Before starting the chapter, it is necessary to explain that using terms such as 'test', 'testing' and 'assessment' interchangeably might be confusing as they, in a technical sense, refer to different but overlapping concepts. As Brown et al. (2002) define, the term 'test' represents any of the various instruments and procedures that are used to gather information about language learners' ability. They argue that assessment, on the other hand, refers to the entire process of gathering information via tests, making interpretations based on the obtained information and arriving at a decision within the language classroom or program (Brown et al., 2002, p. 13). In a broader context of English language teaching, assessment is used to embrace a range of activities and methods of evaluating learners' language ability including teacher evaluation, portfolio, self-assessment and different types of tests. While, testing is often understood as a particular case of evaluating a learner's language ability through a single test administered at one occasion. However, to avoid confusion, the term 'testing' is generally employed throughout this chapter to refer to a single occasion in which the language ability of a test-taker is being assessed or as a general notion of testing. Elsewhere, the term 'assessment' is employed to refer to certain types of testing and testing conditions, e.g. task-based assessment, in a broader context.

3.2 Background to Language Testing

Language tests have always been part of language education, but increasingly they play a more important role in people's lives. McNamara (2000) argues that language tests act " as gateways at important transitional moments in education, in employment, and in moving from one country to another" (p.4). Numerous

educational decisions are made on the basis of the results of language tests. In addition to educational decisions, language tests are used to make decisions that have broader social implications, such as awarding high school diplomas or getting a professional job position. A recent theme in language testing (LT) is the politics of language testing, which examines the hidden agendas and ideologies of the testing industry and of high-stake tests (Shohamy, 2001b; Davis, 2003). Davis (2003) also argues that "tests are inevitably political since what they do – in education as in immigration – is to sort and select to meet society's purpose" (p. 361).

For some years Second Language Acquisition (SLA) and LT were viewed as distinct areas in applied linguistics. The common view was that the two fields did not share common interests and they functioned as two distinct fields of inquiry. The two areas, however, have been in close association with one another since the late 1980s. It is now obvious to researchers from both areas that there are overlapping issues and interests between these sub-divisions of applied linguistics. Carroll (1968) defined the relationship between the two areas by mentioning the important role of a language test as a device to elicit language behaviour which can be studied by SLA researchers. However, it is evident that the relationship between the two goes far beyond what Carroll explained four decades ago. Clearly, language testing both serves and is served by research in language acquisition and language teaching.

In order to have an insight into the early developments of language testing, and as it is needed to provide a background to the discussions in this chapter, a short review of the history of second language testing will be presented. The discussion is aimed at exploring the three distinct traditions that have appeared in LT since the 1950s, which have made dramatic differences in the theoretical and practical aspects of language tests. These are the structuralist, integrative and communicative traditions in LT.

Each tradition will be explained briefly so that the context in which TBA is developed can be better understood.

Before discussing each of the three traditions, it should be noted that, despite all theoretical and practical differences they have, these traditions in LT have developed from the assumptions of a dominant testing theory influencing educational and psychological measurement, i.e. psychometric theory. The psychometric theory of measurement is essentially concerned with describing and measuring characteristics of individuals in a dependable way and is based on a set of assumptions such as the prioritization of psychological processes, construct 'pre-definability' and unidemensionality (Leung, 2003, Leung & Teasdale, 2000). Defining and discussing the principal assumptions of the psychometric theory of measurement is beyond the scope of this chapter. However, it should be noted that the three important traditions of LT discussed here make the same assumptions about the characteristics of individuals, the abilities they test and the relationship between these abilities and test scores.

3.2.1 The Structuralist Tradition

The structuralist tradition within LT draws on the methods used in psychological testing in the first half of the 20th century and on structural linguistics. Psychological tests were developed on the basis of multiple-choice questions, which were recognized as 'objective' measures of testing. From the 1960s, through the 1970s, language testing was essentially informed by a theoretical view of language ability as consisting of language skills, i.e. listening, speaking, reading and writing, as well as the components of grammar, vocabulary and pronunciation (known as the skills and components model). The dominant approach to test design focused on testing

'discrete points' of language, while the primary concern of the field was testing isolated structures and psychometric reliability (e.g. Lado, 1961; Carroll, 1968). For this reason, Spolsky (1977) called it the 'psychometric-structural era'. It was believed that tests which focused on 'discrete' linguistic items were efficient and had favourable reliability of marking associated with objectively scored tests. In order to find out the validity and reliability of tests, it was sufficient to subject test scores to different statistical procedures, such as item and factor analyses.

The structuralist tradition has been criticized on a number of fronts. In fact, its view on language, test design and language construct were all attacked by different researchers at different times (e.g. Oller, 1976; Gipps, 1994). One major criticism of the structuralist tradition to testing was the emphasis it put on reliability and generalizability over a wider notion of construct validity. In other words, for test developers the priority was objectivity and consistency of the test scores and the applicability of the results to a wide range of contexts of language use. All the developments and changes that occurred during the 1960s and 1970s set the context for a new theoretical framework in LT which was later known as the 'integrative tradition'.

3.2.2 The Integrative Tradition

Within the integrative tradition, like the structuralist tradition, language tests prioritize objectivity and reliability and employ the same statistical procedures to ensure these are maintained. In fact, their view of measurement is not much different from that of the structuralist language tests. What is different, however, is the way language is viewed in this tradition. Following Oller (1976), LT research was dominated by the hypothesis that language ability was a single unitary trait. Oller, analyzing the

relationship among scores from a wide variety of language tests, proposed that language proficiency was not composed of discrete elements but consisted of a single unitary ability. The high correlations Oller obtained from different tests made him perceive that the tests were measuring the same factor, which he called 'pragmatic grammar expectancy'. Oller (1976) argued that in order to measure this factor it was necessary to devise tests that investigate the learner's unitary language faculty in holistic activities, such as cloze tests and dictations.

Integrative tests appear to be paying a good deal of attention to validity since they consider language ability to be a unitary construct and they attempt to employ tests that can measure such a unitary construct (the issue of construct validity will be discussed in detail in the following section). Within this framework, it is believed that, because language competence is unitary, it is easy to extrapolate from performance on an integrative test to performance in the real world. Oller's view of language ability, however, has been challenged for both its conceptual and empirical foundations. Ellis (2003), among many other researchers, has argued that the nature of correspondence between the learner's language system, i.e. the 'expectancy grammar', and its use in the context, the 'pragmatic', is not clearly specified and , as a result, leaves the whole idea of 'pragmatic expectancy grammar' vacuous. Hughes and Porter (1983) have questioned the empirical grounds of the factor analyses that Oller used to find empirical support for his findings. They have argued that the use of Principle Component Analysis did not seem to be appropriate because employing such an analysis would make it possible to produce a large first factor whatever the structure of the data.

3.2.3 The Communicative Tradition

During the 1990s both the skills and component model and the unitary trait hypothesis of language ability were crucially criticized. This clearly occurred in the light of the broadened view of language ability, particularly the notions of communicative competence, within the field of language teaching and SLA research. Bachman (2000) contends that the influential views of Widdowson (1978, 1979), Hymes (1972) and Canale and Swain (1980) on language use and language ability as a dynamic and multicomponential construct made a dramatic change in LT. He contends that these views " forced language testers out of their narrow conception of language ability as an isolated trait, and required them to take into consideration the discoursal and sociolinguistic aspects of language use, as well as the context in which it takes place" (Bachman, 2000, p. 3).

McNamara (2000) contends that communicative tests have two features. First, they are performance tests which require assessment to be carried out when the test-taker is engaged in an act of communication. Second, communicative tests pay attention to the social roles test-takers are likely to assume in real world settings.

According to Fulcher (1999), there are three primary aspects of a communicative language test. First, communicative tests involve performance. In effect, the method of testing should ensure that the test performance and the criterion performance are the same. Second, communicative tests are authentic, i.e. the test-takers can recognize the communicative purpose of a task. And finally, communicative tests are scored on real-life outcomes. That is, the essential criterion of success in a language test should be whether the testee performs the task by achieving a satisfactory outcome. Communicative language testing constitutes the fundamental basis for Task-based Language Assessment (TBA) which will be thoroughly examined later in this chapter.

3.3 Key Properties of Language Tests

The fundamental purpose of a language test is to present a measure that can be interpreted as an indicator of an individual's language ability. A good language test, therefore, is supposed to provide a true measure of a test-taker's language ability. Increasingly, language testers are concerned with the usefulness of language tests. Bachman and Palmer (1996) describe usefulness as a function of different properties such as reliability, construct validity, authenticity, interactiveness, impact and practicality. However, the most substantial of these properties, particularly as they relate to the purpose of this research, are validity, reliability and authenticity of a test. These properties continue to affect all aspects of test design, test use, and interpretation of test results. Various issues relating to each of the above-mentioned properties of a test would require extensive discussions, which would go beyond the scope of the present discussion. Hence, only a brief account of each property would be provided here.

3.3.1 Validity

Validity primarily deals with the concept of whether a test measures what it purports to measure. By definition, validity refers to the extent to which the inferences or decisions that are made on the basis of test scores are meaningful, appropriate and useful. Traditionally, testers have distinguished a number of different types of validity: content, predictive, concurrent, construct and face validity (Davis, 1978; Hughes, 1989). It was believed that validity was a particular characteristic of a test, which consisted of different components with different values. Therefore, a test was likely to have one type of validity but lack the others. Messick (1989, 1994, 1996) has challenged this perspective to validity and argued that construct validity is a

multifaceted but coherent and overarching concept. Messick (1989) describes validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p.13). He has argued that validity is not a characteristic of a test, but a feature of the inferences made on the basis of test scores and the uses to which a test is put. In fact, it is not the test that is being validated but the principles for making inferences. This concern about test score interpretations and uses has inevitably raised the issue of test consequences, or, as it is known the consequential validity of a test. Messick (1989) proposes a framework, which he calls a 'progressive matrix', for validity as a unitary concept (Figure 3.1).

Figure: 3.1

Messick's Progressive Matrix

	Inferences	Uses
Evidence	Construct validity	Construct validity + Relevance/utility
Consequences	Construct validity + Value implications	Construct validity + Value implications + Relevance/utility + Social consequences

As indicated in the model, the columns represent the outcomes of testing and the rows represent the types of arguments that should be used to justify testing outcomes. Each of the cells contains 'construct validity', but new facets are added as one goes through the matrix from top left to bottom right. Thus, for justifying a particular interpretation of a test score, testers should gather evidence for construct validity and consider the value implications of this interpretation. If the test score is being used for a particular

purpose, justifications must be made by considering not only the construct validity and value implications, but also the relevance or utility of the particular use and the social consequences of using the test score in a particular way.

As influenced by Messick's unified perspective of validity, validation is now seen as an ongoing process of continuous monitoring and updating of relevant information that is never complete. Validation is known as the empirical evaluation of the meaning and consequences of measurement. In a later section on task-based assessment, the validity of tasks that are employed in the present research to assess oral language ability will be carefully discussed.

3.3.2 Reliability

'Reliability' is often defined as the consistency of measures. Reliability is a quality of test scores, which inquires whether the score is free from errors of measurement. There are many factors other than the ability being measured that can affect performance on tests and would therefore constitute sources of measurement error. Differences in testing condition, fatigue, anxiety, and other similar factors may contribute to measurement error. Reliability, in effect, refers to the consistency of measures across different times, test forms, raters, and other characteristics of the measurement context.

Traditionally, reliability was measured directly in various ways that can be generalized to a comparison between one set of items and a comparable set in order to estimate consistency of measure. Different methods of parallel tests, split-half, and test-retest are the most frequently used measures of estimating reliability (Alderson & Banerjee, 2003b). More recently, reliability is measured through Generalizability theory (G-Theory) or Item Response Theory (IRT) (Bachman, 1990). G-Theory is

based on factorial design and analysis of variance and enables testers to estimate the effects of multiple sources of measurement error. IRT helps testers to estimate the statistical properties of items and the abilities of test takers independently of a particular group of test takers or a particular form of a test (Bachman, 1990).

Although two separate definitions can be conveniently given while defining the concepts of reliability and validity, practically, these are two interrelated and overlapping qualities of a test. Davies (1978) argues that if the reliability of a test is maximized, it might be at the expense of validity, and if validity is maximized, it is likely to be at the expense of reliability. Research in LT has revealed that validity and reliability are two complementary qualities of a test, since a test needs to be reliable to be valid. In other words, reliability is a necessary, but not sufficient, condition for construct validity.

Alderson (1991) problematizes the distinction between reliability and validity. He argues that, although the difference between the two is in theory clear, problems arise when considering how reliability is measured. Alderson (1991) points out that test-retest reliability is the easiest measure of reliability to conceptualize. However, there are problems with this concept. In theory, if a person takes the same test on a second occasion, and the test is reliable, the score should remain constant. But the score might have changed probably because test-takers have learned from the first administration or because their ability has changed. In this case a lower test-retest correlation might be observed, and this would be a valid indication of change in ability but an indication of lack of reliability at the same time.

Parallel tests are another measure of reliability. But parallel forms of a test are often validated by correlations (concurrent validity), and so high correlations between two parallel forms would be a measure of validity and not reliability. Another argument

is put forward as SLA research shows that learners vary in their performance on different tasks, and that this variation can be systematic (Swain, 1993, among many other researchers). This systematic variance seems to be present in second language performance. Therefore, a low reliability coefficient is expected on two performances of the same individual on the same task. It can be concluded that in the light of what research has indicated about variation in language performance, the way reliability is conceptualized and operationalized is problematic. Some researchers argue that, given Messick's unitary view of validity, reliability is conceptually merging into a unified view of validity (Alderson and Banerjee, 2003b).

The two concepts of reliability and validity are constantly considered important qualities of a language test. However, it should be noted that reliability and validity are inherently related to the use of tests for various purposes, and are not inherent qualities of the test itself. In other words, as Brown et al. (2002) contend "it is probably useful to think of reliability and validity as processes, that is, the processes of gathering evidence about the particular use of a test" (p. 14). In addition to reliability and validity, authenticity is another important quality of a test which has received considerable attention since the communicative tradition in LT emerged. In the next section, the authenticity of language tests will be introduced and discussed.

3.3.3 Authenticity

Since the advent of communicative language testing in the 1970s, authenticity has been a great concern in language testing. It has often been argued that, to be able to predict a test-taker's ability to communicate in the real world, tests should be as similar to that real world as possible. LT research acknowledges that authenticity should be considered as a critical quality of language tests, alongside validity and

reliability (Alderson and Banerjee, 2003a; Bachman and Palmer, 1996; Messick, 1996). Bachman (1990, 1991) builds on Widdowson's (1978) definition of authenticity and describes it as "appropriateness of the language user's response to language as a means of communication" (p.304). Bachman and Palmer (1996) define authenticity as " the degree of correspondence of the characteristics of a given language test task to features of a TLU [target language use] task" (p.23). As Chalhoub-Deville (2001) puts it, authenticity is "the establishment of a more direct relationship between language use and activities employed in instruction and assessment" (p. 216).

Although there is general consensus that authenticity is an important property of a test, there are different arguments about the criteria for measuring authenticity. First, it is true that materials and tasks in language tests can be relatively realistic but they can never be thoroughly real. In effect, authenticity is a relative quality with some tasks being more authentic than others. In other words, as Lewkowicz (2000) argues "tasks would not necessarily be either authentic or inauthentic but would lie on a continuum which would be determined by the extent to which the assessment task related to the context in which it would be normally performed in real life" (p.48). Therefore, authenticity is very much dependent on the degree to which test materials and conditions replicate real life situations. It is also argued that authenticity is an important quality of language tests because it is closely related to how language ability is defined and how the results of language tests are interpreted (Bachman, 1990). Finally, a test can never be entirely authentic because test performance does not exist for its own sake and the test-taker is aware of this.

As discussed in previous chapters, a major part of the present research deals with assessing language ability. It is necessary to know the existing language ability

models and discuss a dominant model which is accepted by task-based approaches to language instruction and assessment. Hence, in the next section, prominent language ability models are introduces and evaluated.

3.4 Models of Language Ability

It is apparent that a clear and explicit theory of language ability is essential to language test development and use. Models of language ability, normally inspired by language teaching and SLA research, would help formulate and develop a theory of performance. In effect, by the advent of communicative language teaching and testing, research in language testing (LT) and language teaching has been convinced that a model of underlying capacities in performance, other than linguistic knowledge, is necessary if language testing is hoping to approach performance testing. McNamara (1995) in an article 'Modelling performance: Opening Pandora's box' insists that a model of abilities in performance is required and will help solve the problem of generalizing from one observed instance of behaviour to other unobserved instances. He adds that a performance model will help researchers and language testers understand the role of non-linguistic factors and the performance of native speakers. Eventually, a model is required to provide a theory that informs the research agenda about the role of non-language specific cognitive and affective variables in language performance settings. This model should also present a general framework within which explicit hypotheses can be formulated about the relationship between test-taker and rater behaviour and test score.

A primary model of language ability proposed by Hymes (1972) has been the first theory of language performance available to the communicative testing tradition. As discussed in Chapter II, Hymes started defining the notions of communicative

competence and distinguished between actual instances of language use in real time and abstract models of underlying knowledge and capacities involved in language use (See Chapter II for a detailed discussion). Based on Hymes's idea of communicative competence, Canale and Swain (1980) developed a model of language performance. The model consisted of grammatical competence, strategic competence and sociolinguistic competence. Examining the theoretical bases of language teaching and language testing, Canale and Swain (1980) distinguished 'grammatical competence', which includes lexis, morphology, sentence grammar, semantics and phonology, from 'sociolinguistic competence', which consists of sociocultural rules. By strategic competence, they referred to possession of 'coping strategies' in actual performance in the face of inadequacies in any other areas of competence. Canale (1983) added a new category of competence by making a further distinction between sociolinguistic competence and discourse competence, which refers to mastery of how to combine grammatical forms and meanings to achieve a unified spoken or written text in different genres. He proposed that discourse competence would help unity of a text be achieved through cohesion in form and coherence in meaning. The original feature of this model was the theorizing of the domains of language knowledge to include sociolinguistic competence and other areas of competence.

Although Canale and Swain's model was dominant for a decade, it was later criticized for some of its principles. Canale and Swain (1980) argue that while performance may demonstrate such factors as volition and motivation, they "doubt that there is any theory of human action that can adequately explicate ability for use" (p.7). They argue that, as this ability cannot be modeled it cannot be included in their framework. McNamara (1996) argues that Canale and Swain have simply chosen to exclude 'ability for use' from their definition of communicative competence. This means that

their model lacks a notion of potential for use or underlying skill. Canale and Swain (1980) view 'ability for use' as simply part of what they call 'communicative performance', which they define as "the realization of these competencies [i.e. the components of communicative competence proposed in their model] and their interaction in the actual production and comprehension of utterances" (p.6).

A second criticism of this model concerns the notions of strategic competence as put forward by Canale and Swain. In the definition of strategic competence, they do not clearly discuss whether the strategies employed by L2 users are gained as a skill or acquired in the form of knowledge. McNamara (1995) argues that " It is hard to see that what is involved here is knowledge rather than ability or skill" (p.168). Another problematic feature of Canale and Swain model is that the interaction between the components has been largely ignored. However, as Canale (1983) acknowledges "this theoretical framework is not a model of communicative competence, where model implies some specification of the manner and order in which the components interact and in which the various competencies are normally acquired" (p.12). In other words, Canale and Swain's model was not intended to account for the way the components may interact with each other.

Bachman (1990) proposes a model of 'communicative language ability', which is considered the most comprehensive model of language performance in LT. Primarily, Bachman defines communicative language ability as consisting of both knowledge, i.e. competence, and the capacity for implementing or executing that competence in appropriate contextualized communicative language use. The framework he proposes includes three components: language competence, strategic competence and psycho-physiological components that are used in communication via language. Figure 3.2 shows Bachman's proposed model of language ability (Bachman, 1990, p. 84).

Figure 3.2

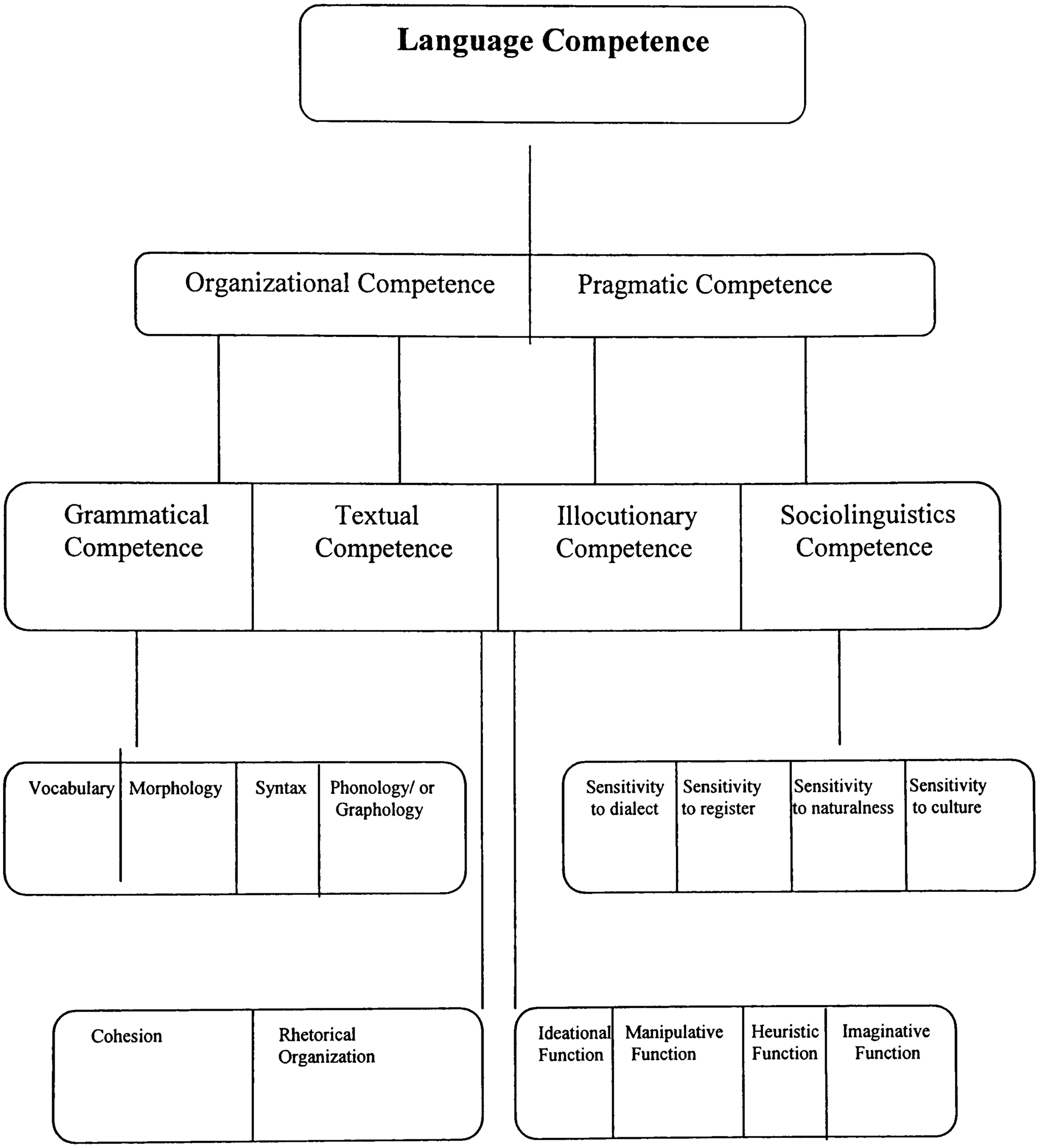
Bachman's (1990) Model of Communicative Language Ability

While Figure 3.2 shows different competencies in communicative language ability, Figure 3.3 demonstrates the components of language competence as proposed by Bachman (1990). He views language competence as mainly consisting of organizational competence and pragmatic competence, each with a number of sub-competencies.

Bachman's (1990) organizational competence includes grammatical competence and textual competence, whereas his pragmatic competence consists of illocutionary competence as well as sociolinguistic competence. Grammatical competence, as demonstrated in Figure 3.3, includes those competencies that are involved in language usage. These consist of a number of relatively independent competencies such as knowledge of vocabulary, morphology, syntax, and phonology and/or graphology. On the other hand, textual competence includes knowledge of the conventions for joining utterances together to form a text, which is essentially a unit of spoken or written language, consisting of two or more utterances that are structured according to rules of cohesion and rhetorical organization (Bachman, 1990, p. 87).

Figure 3.3

Bachman's (1990) Model of Language Competence



Pragmatic competence in Bachman's model originates from Van Dijk's (1977) description of pragmatics. Bachman, following Van Dijk, views pragmatics as being concerned with the relationship between utterances and the acts or functions that speakers or writers intend to perform through utterances. This function is referred to as the illocutionary force of utterances, and refers to the characteristics of the context

of language use that determine the appropriateness of utterances. In his model, Bachman presents pragmatic competence that includes both illocutionary competence and sociolinguistic competence. His illocutionary competence includes a number of language functions such as ideational, manipulative, heuristic, and imaginative functions of language.

In Bachman's model, illocutionary competence explains how learners are able to use the language to express a wide range of functions and to interpret the illocutionary force of discourse. Sociolinguistic competence deals with the appropriateness of those functions in the context of language in use. In other words, sociolinguistic competence, in Bachman's model, is the sensitivity to, or control of the conventions of language use that are determined by the features of the specific language use context. Sociolinguistic competence, in fact, enables speakers to perform language functions in ways that are appropriate to the context. Finally, sociolinguistic competence includes sensitivity to differences in dialect or variety, sensitivity to registers, sensitivity to naturalness, and the ability to interpret cultural references and figures of speech.

A second type of competence that Bachman proposes in his model is strategic competence. Bachman views strategic competence as an important part of all communicative language use, not just that in which language abilities are ineffective and must be compensated for by other means. It is worth mentioning that to Bachman (1990) "strategic competence is not part of language competence" (p. 106). He assumes strategic competence as a general ability, which enables an individual to make the most effective use of available abilities in carrying out a given task. For example, in a test of reading comprehension, answering questions that require inferences needs strategic competence, in that the test-taker must recognize what

information outside the discourse itself is relevant to answering the question, and then must search for that information in his memory. Therefore, it is understood here that Bachman views strategic competence more as an ability, capability or capacity than an area of knowledge. In this model, strategic competence includes three components of assessment, planning and execution.

Finally, the psycho-physiological competence, in Bachman's model, refers to mechanisms that are essentially the neurological and physiological processes involved in the execution phase of language use. For instance, the visual and audio channels or productive and receptive modes through which language is processed are distinguished and have to be taken into consideration while assessing communicative language ability.

To date, Bachman's model has proved to be the most comprehensive performance models in LT. It appears to be more adequate than the model presented by Canale and Swain (1980), in terms of the detailed specifications of the component parts. One important feature in Bachman's model is the elaborate description of language competence it presents. Moreover, the separation of strategic competence from language competence in this model is a significant improvement. The model helps clarify our conceptualization of language performance in test settings, and enables investigations of the claims made by tests that are assessing communicative language ability. However, a number of criticisms have been put forward regarding different aspects of this model. McNamara (1995) argues that in Bachman's model "it is not clear to what extent strategic competence includes general affective factors as well as cognitive ones" (p. 171). Bachman's definition of illocutionary competence has also been criticized. Like Canale's notions of discourse competence, Bachman's definitions of illocutionary competence seem to be unclear. In certain places he

seems to be suggesting that illocutionary competence is part of strategic competence (Bachman, 1990, p. 104). McNamara (1995) argues that if illocutionary competence were taken to include only routinized realizations of language functions, then illocutionary competence could be seen as a form of knowledge.

More recently, Chalhoub-Deville (2003) has re-investigated Bachman's (1990) model of communicative language ability from a new perspective. She argues that' although Bachman has considered an interactional perspective in his model, this interaction is incorporated from an individual-focused cognitive perspective. In fact, she points out that the present communicative language ability models are individual focused and are largely a representation of cognitive or 'within-language user' constructs, while an interactional approach to language ability views L2 construct as socially and culturally mediated. She proposes that an alternative view in which individual ability and contextual facets interact in ways that can change them both is needed. She suggests that 'an ability – in language user – in context' view is to be preferred over Bachman's 'ability – in language user' presentation. What she insists on is consideration of a social interactional perspective in models of language performance.

Skehan (1998, 2001) has also criticized ability oriented proficiency models (Canale and Swain, 1980; Bachman, 1990). He proposes a model of oral test performance which provides a more appropriate context for assessing oral language ability in TBA. Skehan's model will be explained and discussed in Section 3.7 on Task-Based Assessment later in the current chapter.

Although Bachman's model has been criticized for some of its assumptions and principles, it is still the most prominent model of language ability known to LT research. This model is usually adopted to represent general language ability in LT research. However, in the context of TBA, all previous language models appear to be

insufficient in terms of indicating the probable influence of task characteristics on language performance. These models of language performance have failed to consider the potentiality of interaction between task characteristics, interactive conditions and a test-taker's performance on task. Later in this chapter, I will present and discuss Skehan's proposed model of oral language performance which takes the drawbacks of the previous models into account and attempts to provide a more appropriate model to be used in task-based assessment.

As oral language ability is being specifically assessed through tasks in the present research, I will first explain the construct of oral ability and investigate the common tests that are frequently employed to assess oral language ability.

3.5 Oral Language Ability

In this section, I will try to provide a definition for oral language ability as this ability will be observed and assessed in both studies reported in this research. Oral language ability, like other language abilities, is usually assessed by utilizing tests that evaluate a test-taker's language performance in a defined context. However, what this oral ability refers to is a challenge LT has to deal with.

An ability that is defined for the purpose of measurement is normally called a 'construct'. Carroll (1987) defined a 'construct' of mental ability in terms of a particular set of mental tasks that an individual is required to perform on a given test. Cronbach and Meehl (1955) defined a construct as a " postulated attribute of people, assumed to be reflected in test performance" (p. 283). Chappelle (1998) views a 'construct' as a meaningful interpretation of an individual's behaviour. She distinguishes among three perspectives on construct definition: a construct may be defined as a trait, as a behaviour, or as some combination of trait and behaviour. In a

trait definition of a construct, as Chappelle argues, a person's consistent performance on a test is taken to be a sign of fairly stable configuration of knowledge and skills that the person possesses and can apply in all contexts. The trait theory position assumes that the test scores are not task specific, or the tasks are for the most part interchangeable. Hence, the scores represent underlying constructs that enable speech, and from which we can generalize to other speaking tasks in other tests.

In contrast, in defining construct as behaviour, it is believed that test variations are very common both due to task features and to learners' variations. In fact, test performance is assumed to show the results of an individual's performance on a specific task or in a specific context, but not on other tasks or other contexts. Therefore, the inferences from scores may only be generalized to identical tasks in other tests or the real world (Fulcher, 2003).

Young (2000) argues that neither trait nor behaviour definitions are satisfactory for theories of language in use, such as communicative competence. Drawing upon Bachman's (1990) definition of communicative language ability, Young (2000) proposes that definitions of a construct can only be acceptable when they are based on both a trait and behaviour. He argues that this is inevitable because the communicative ability is itself based on both knowledge and the capacity for implementing or executing that competence in a specific context of use. Therefore, oral language ability, in this research, is considered as a combination of both a trait and behaviour, i.e. both the knowledge and the capacity of implementing the knowledge in actual situations. The oral language tests employed in this research, i.e. narrative tasks, are believed to be eliciting samples of an individual language knowledge and the capacity for implementing such knowledge in context of authentic

language use. In the following section, I attempt to show how oral language ability has been assessed and how oral language tests have changed during the past decades.

3.6 Oral Language Tests

The testing of speaking has a relatively long history, but it was not until the 1980s, following the development of communicative language teaching, that the direct testing of L2 oral proficiency became commonplace. Oral interviews, such as the ones developed by Foreign Service Institute (FSI) and Oral Proficiency Interviews (OPI), were long considered as valid direct tests of speaking ability. The OPI is modeled after the (FSI) oral interview in its structure, rating criteria and level descriptors. Chalhoub-Deville (2001) defines OPI as "a structured, live conversation between a trained interlocutor/rater and a test-taker on a series of topics of varied language difficulty, with the goal of establishing the test-taker's proficiency level" (p. 213). The interviewer initiates the interactions and builds on the responses of the interviewee. The testers use a set of guidelines for scoring the interview.

OPIs are still in use all over the world. However, these oral interviews have been recently criticized for not possessing some of the key properties of language tests. A number of research studies have been carried out to evaluate different aspects of these oral interviews. Research in the field of discourse and conversational analyses have clearly demonstrated that oral interviews are only one of many possible genres of oral test tasks. It is also evident that the language elicited by OPI is not the same as that elicited by other types of tasks. Some researchers have expressed doubt about the capacity of oral tests to sample sufficient language for accurate judgements of proficiency (Hall, 1993, Norris, 1991).

Following the frequent criticism researchers expressed on the consistency of OPI rating guidelines, the American Council on the Teaching of Foreign Languages (the ACTFL guidelines) published an influential set of guidelines for the assessment of oral language proficiency in 1986. These guidelines include nine-level descriptions rating from the novice to the superior, explaining proficiency of each level in detail. This was followed by the introduction of the widely influential ACTFL Oral Proficiency Interview (ACTFL OPI).

A growing body of research has attempted to investigate different aspects of both ACTFL OPI for issues of construct validity (Henning, 1990), validity of scores and rating scales (Reed, 1992), and rater behaviour and performance (Thompson, 1995). Conclusions have varied, with some researchers arguing for the usefulness and validity of the OPI and its accompanying rating scales, and others criticizing the tests. Fulcher (1997) argues that speaking tests are particularly problematic from the point of view of reliability, validity, practicality and generalizabilty, which are also the underlying debates about ACTFL and OPI.

During the past decade, as the common oral language tests have been criticized on the grounds of validity, reliability and authenticity, task-based oral assessment has become a dominant approach to assessing oral language ability. In this approach, a number of different tasks are normally employed to elicit sufficient samples of test-takers' language performance to represent different relevant competencies and skills involved in oral ability. Through a number of tasks, samples of monologic, dialogic and interactive oral language performance of test-takers are elicited. A test-taker's language performance is then recorded and rated by at least two trained raters, on a large number of linguistic, discoursal, sociolinguistic, and communicative criteria.

The detailed discussions of assessing oral language ability in TBA will be explained in the following section.

3.7 Task-Based Assessment (TBA)

The primary purpose of this section is to explore what 'tasks' are in LT and how they are employed in TBA to assess learners' communicative ability in a second language. Definitions and different types of TBA will be then evaluated. How performance on task is measured is an issue of controversy in TBA, which will be discussed in this section. The last part of this discussion will deal with problems of TBA.

Since discussions of TBA are continuously connected to or built upon issues of performance testing, it is necessary to explain performance testing before dealing with any definitions of tasks or TBA. As language instruction has, during the past two decades, focused on educational outcomes in terms of second language use, language assessment has likewise started focusing on evaluating what learners can do with the language. To achieve this purpose, language testing research has recently paid considerable attention to the development of various approaches to second language performance assessment (e.g. McNamara, 1996; Messick, 1996; Norris et al. 1998). LT researchers (e.g. Carroll, 1985; Clark, 1975; Henning, 1987) initially started using terms such as 'direct' and indirect' tests to show the extent to which testing formats and procedures attempt to duplicate as closely as possible the setting and operation of the real-life situations in which the proficiency is normally demonstrated (as discussed in section 3.3.3). Following the same approach, other researchers (e.g. Bailey, 1985; McNamara, 1996; Wesche, 1985) have used the term 'performance test' and 'performance assessment' to characterize measurement procedures that approximate non-test language performance. Bachman (2002) defines it in a general context by

contending that "performance assessments are typically designed to assess complex abilities that cannot easily be defined in terms of a single trait, and typically present test-takers with tasks that are much more complex than traditional constructed-response items" (p. 471). This definition is helpful as it refers to both theoretical assumptions of performance testing and to the practical conditions of such tests. As Bachman argues, performance testing has become significant since language testers have realized that language ability is not a pure 'trait construct'. Moreover, his definition indicates that performance assessment is more complex and more advanced compared to traditional tests of language structure.

At first, tests were recognized as being either performance or non-performance (Wesche, 1985). However, later language testers (Norris et al., 1998) acknowledged that it is more appropriate to consider performance testing along a continuum from least direct and least real-world (or least authentic) to most direct and most real-world (or most authentic). In effect, all language tests are known to have some degree of performance included. The most significant advantages of performance assessment, as mentioned by LT researchers, are: (1) performance assessment would compensate for the negative washback effect and the limited content coverage of standard testing; (2) it is more valid and authentic than non-performance tests; (3) it approximates the conditions of real-life, and (4) it can predict students' ability in future (Brown et al., (2002; Shohamy, 1995).

3.7.1 Tasks and TBA

During the 1990s, following the innovations of communicative and task-based language teaching, task-based approaches to assessment became very popular. Language testers decided to develop various forms of assessment which were aimed

at providing information on how well learners were able to mobilize language to achieve meaningful communicative goals. To achieve this objective, they realized that they needed tasks to elicit performances that could be used not only to assess language ability but also to evaluate whether the test-taker could perform some specific real-world activities.

Detailed definitions and discussions of 'task' and 'task difficulty', in the context of instruction, were provided in Chapter II. Although tasks in TBA originally represent the concept of task in task-based instruction, language-testing perspective on task should also be taken into consideration here. Before defining a task in TBA, it should be noted that language testers have used the term 'task' variably. The term was initially used to refer to any devices employed for assessing language ability (Chalhoub-Deville, 2001). In this sense, a multiple-choice item of grammar or a free composition is as much a 'task' as an information gap activity or an oral interview. Ellis (2003) argues that this refers to Breen's (1987) broad notion of tasks in language pedagogy. More recently, in the context of performance assessment, task assumes a narrower meaning (which was discussed in detail in Chapter II). In this narrower sense, assessment tasks are viewed as devices for eliciting a test-taker's communicative performance in the context of language use that is meaning-focused and directed towards some specific goals (Ellis, 2003). Task, in the current discussions of task-based assessment, then refers to this latter concept.

Although clear definitions exist in terms of the distinction between tasks and non-task activities, the actual distinction between the two is not always as clear. Baker (1990) and Ellis (2003), among others, argue that there exists a continuum rather than a dichotomy. It will not always be easy to determine whether a particular test is task-based. Many instances can be mentioned in which tests are located along this

continuum. For example, a listening comprehension test where test-takers are asked to listen to a contrived mini-lecture and then answer a number of multiple-choice questions to demonstrate their comprehension is a good example of such a test.

Ellis (2003) argues that just as language teaching methodologies believe that tasks constitute *the prima facie* means for promoting acquisition of an L2, so language testers have increasingly recognized the value of tasks for assessing learners' capacity to communicate in an L2. McNamara (1996) notes that performance tests based on tasks have arisen both because of the need to develop selection procedures for specific groups of L2 learners and the need to bring testing in line with the developments in language teaching. Brindley (1994) identifies a number of features of what he calls 'Task-centered assessment'. He believes that task-centered assessment results in both teachers and learners focusing on language as a tool; it enables assessment to be more easily integrated into the learning process; it provides learners with useful diagnostic feedback on progress and achievement; and it enables the results of an assessment to be reported in a way that is intelligible to non-specialists. Brindley (1994) proposes one of the basic definitions of task-based assessment:

> Task-centered language assessment is the process of evaluating, in relation to a set of explicitly stated criteria, the quality of the communicative performances elicited from learners as part of the goal-directed, meaning-focused language use requiring the integration of skills and knowledge. (p. 77)

Brindley, as it is understood from his definition of task-based language assessment, views language proficiency as encompassing both 'knowledge' and 'ability for use'. It should be noted that in his elaborate definition, however, Brindley considers task assessment from a classroom perspective which might be different from task-based assessment in an assessment setting.

Chalhoub-Deville (2001), on the other hand, considers oral language tasks from a testing perspective and identifies three key characteristics for TBA. According to her, TBA must reflect learner-centered properties; that is, the tasks must not be 'conformity-oriented' or 'practice-oriented' but must encourage individual expression. In effect, a learner-centered assessment gives the test-takers an opportunity to utilize their background knowledge and experience to be active and autonomous in the communicative activity they are engaged in. Second, performance on tasks must be contextualized, which can only be achieved by using 'meaningful situations' and requiring 'extended discourse'. Third, tasks should be authentic in terms of real-life use, i.e. they should mirror as closely as possible target language use tasks. Chalhoub-Deville (2001) criticizes the current approaches to TBA because they are unable to uncover the specific language abilities underlying performance. She further calls for an approach to TBA in which test specifications would include the knowledge and skills that underlie the language construct (Chalhoub-Deville, 2001, p. 225).

A different approach to defining TBA can be found in the work of a number of researchers such as Norris et al. (1998), Long and Norris (2000) and Brown et al. (2002). Long and Norris (2000) distinguish task-based from other forms of language performance assessment as follows:

> [T]ask-based language assessment takes the task itself as the fundamental unit of analysis motivating item selection, test instrument construction, and the rating of task performance. Task-based assessment does not simply utilize the real-world task as a means for eliciting particular components of the language system, which are then measured or evaluated; instead, the construct of interest is performance of the task itself. (p. 60)

As stated in the definition, proponents of this approach to TBA are interested in eliciting and evaluating learners' abilities to accomplish particular tasks in which target language communication is essential (Brown et al., 2002). In effect, they focus on the actual relationship between task features and the behaviours they elicit, on investigating how tasks are actually accomplished and on understanding what makes a given task more or less difficult for different examinees (Norris et al., 2002). Norris et al. (1998) suggest that "L2 performance assessment should encompass evaluation of learner performance within the range of task-inherent ability requirements as well as task characteristics that are found in a given task" (Norris et al., 1998, p. 56). In this approach, like other approaches to TBA, interpretations about task performance are ideally based on the criteria associated with real-world expectations for task accomplishment. What distinguishes this approach from others, however, is the emphasis it puts on whether the language learners are able to accomplish the task according to real-world criteria. This emphasis, in effect, implies that what seems to be significant in this approach is the test-takers' 'ability' to accomplish particular tasks or task types. In taking this approach, Brown et al. (2002) appear to be defining the 'construct' in terms of what test-takers can do, which would limit their interpretations and predictions about a test-taker future performance. In other words, it appears that their definition of construct is limited to a behaviorist approach in which 'trait' has not been clearly defined (See section 3.6 for a detailed discussion of construct).

The discussions of TBA evidently indicate that there is common consent about understanding of tasks and task-based assessment among many LT and SLA researchers. However, not much agreement could be reached among LT researchers in terms of what the 'construct' of language ability is or how to measure language

performance in TBA. At least three types of TBA exist, which view measurement of language performance in a different way. A brief discussion of these types of TBA follows.

3.7.2 Types of TBA

There are several types of task-based assessment, each tied to a range of decision – making purposes. Each of these approaches incorporates specific procedures for analyzing and evaluating task performance. One approach involves the assessment of task outcomes in terms of a learner's failure or success to accomplish the task. Such an approach would be utilized when it is necessary to certify that a learner is able to accomplish a set of tasks identified in a curriculum. Outcome-referenced tests are criticized because results from such tests cannot be conveniently generalized across programs since the success or failure in performing a task would not say much about the details of the language ability which is being tested (Robinson, 1996). Nor could they inform teachers and testers about the efficiency with which the learner performs the task (Norris, et. al., 1998).

Robinson (1996) distinguishes between outcome-referenced and system-referenced tests. Drawing on Baker (1990), Robinson indicates that system-referenced tests attempt to tap a particular psychological construct, which underlies a particular language task, without analyzing the accomplishment of the task itself. In fact, such tests assess knowledge of language proficiency in a general sense without reference to any particular use or situation. Thus, tests that belong to structuralist and integrative traditions and even some of those belonging to the communicative tradition are system-referenced.

Performance-referenced tests, on the other hand, attempt to provide information about the ability to use the language in specific contexts. They are directed at assessing a particular performance like making a hotel reservation on the phone or giving directions to someone to find an address. Robinson (1996) contends that performance-referenced tests approximate as closely as possible the conditions of a future language task, and they therefore retain high validity. The defining feature of such a test is that it is performance rather than system which is being assessed.

For example, in a direct performance-referenced test a doctor's oral language ability is assessed in the real communication he or she has with patients while examining them. Performance-referenced assessment is the most common type of TBA employed by a large number of researchers (Brown et al., 2002; Ellis, 2003; Skehan and Foster, 1997; Wigglesworth, 1997, 2001). It is also employed by a number of international language testing organizations, e.g. TOEFL, TSE (Test of Spoken English), and YLE (Young Learner English exam). In the section that follows, I will explain a model of oral language ability which has carefully considered the significant aspects of performance-referenced assessment in TBA. This model is increasingly used and recognized as a well-developed framework for the assessment of oral ability.

3.7.3 Skehan's Model of Oral Language Performance

As discussed before, LT researchers have repeatedly criticized the existing models of language ability. Skehan (1998) reasons that predominant approaches to language testing (e.g., Bachman, 1990; Bachman and Palmer, 1996) have overemphasized the search for an underlying 'structure-of-abilities' that L2 learners acquire. He contends that the recent move towards tasks has posed problems for abilities-oriented proficiency models of testing, i.e. Canale and Swain (1980) and Bachman (1990). He

argues that these models posit an underlying structure of competence, and then propose mediating mechanisms by which such competence will impact upon performance. He states that:

> In principle, such an approach might be extremely rewarding but, in practice, the codifying nature of the underlying competence-oriented models has not interfaced easily with effective predictions to real world performances (Harley et al., 1990; Skehan, 1998). At the most general level, the problem is that underlying and generalized competences do not easily predict across different performance conditions and contexts. Moving from underlying constructs to actual language use has proved problematic. (Skehan, 2001, p. 167)

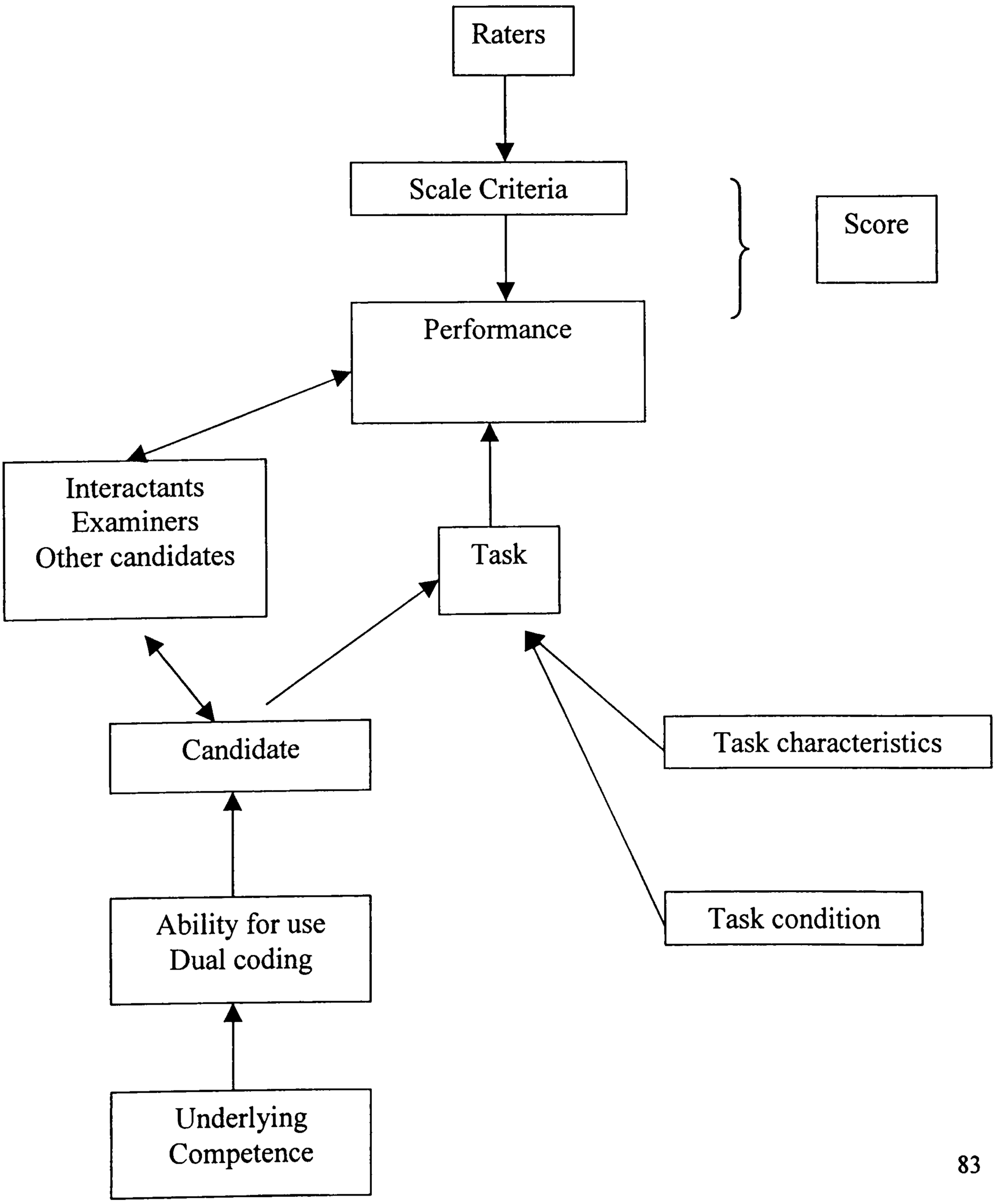
Skehan (1998, 2001) draws upon Kenyon (1992) and McNamara (1995) and proposes a model of oral language performance. He states that all these models attempt to portray the assessment event in more comprehensive ways which (a) incorporate a large number of performance elements directly, and (b) clarify how research studies might be organized and integrated more effectively to give an empirical basis for the claims that are made about spoken language assessment. His model is shown in figure 3.4. The model in general clarifies the potential fallibility of a test score as an indicator of language underlying abilities. As indicated in the model, a test score is most immediately influenced by the rating procedures and rating scales that are being used, as well as by the raters who are judging the performance. In addition, the performance that is being rated will be filtered through a rating scale. These rating scales may vary in their origin, their characteristics and in their purposes. Skehan (2001) proposes that, "as a result of these rater and scales factors, we have to consider the possibility that the score assigned to a test-taker may not reflect his/her

performance only, but may partly be based on basis and limitations arising from raters and scales" (p.168).

A number of additional influences on performance are recognized in this model. For instance, the interactive conditions under which performance is elicited, the relevant abilities of the test-takers, the tasks that are used to generate the performance and the conditions under which the task is completed could all influence the oral performance of a test-taker.

Figure 3.4

Skehan's Model of Oral Language Performance



Skehan (1998) suggests that as an alternative it is essential to investigate performance and processing in their own right, because "these factors are fundamental for generalizations that need to be made about how language will actually be used" (p. 264). He goes on to note that tasks themselves act on and influence an individual's L2 performance and also judgments or ratings of their L2 proficiency. Therefore, it is necessary that tasks are carefully analyzed so that a better understanding of how they affect a given performance is achieved. He argues that, based on evidence from studies of the influence of task factors on L2 performances, the processing dimensions noted in Chapter II, e.g. cognitive complexity, can serve as useful indices for task difficulty analysis. He believes that sampling of L2 performance on tasks would necessarily come out of an understanding of the processing attributes that are inherent in real-world tasks selected from a needs analysis.

As mentioned in the previous section, this model of language ability has increasingly attracted attention among task-based researchers (Chalhoub-Deville, 2001; Wigglesworth, 2001). Within a cognitive approach to task-based research, this model clearly demonstrates that task characteristics and performance conditions, among other factors, would impact on language performance on the task. The model appreciates the potential effects of cognitive processes on performance, considers the influence of raters and rating scales on performance, and provides a practical framework for assessing language performance on tasks.

3.7.4 Measuring Performance in TBA

It is obvious that tasks do not themselves provide a measure of test-takers' language ability. Tasks are essentially used to elicit a performance, which then needs to be assessed in order to provide a clear measure of a test-taker's language ability.

Therefore, a crucial dimension of TBA is the method employed to assess language performance. Ellis (2003) mentions four principle methods for measuring performance in TBA: direct assessment of task outcome, external rating procedures, analytic detailed measures, and self assessment. The two methods which, for reasons of relevance and context, are frequently used in LT research are the rating procedures and analytic detailed measures.

3.7.4.1 Rating Procedures. Rating procedures are the most common method of assessing language performance both in the system-referenced and performance-referenced testing. Employing rating scales and raters, specifying competency and levels of performance, and rater training are the significant features of rating procedures. As McNamara (2000) explains, in rating procedures there is an agreement about the conditions (including the length of time) under which a test-taker is performing; then certain features of the performance, e.g. fluency, accuracy, organization, sociocultural appropriateness, are judged; the weighting and relevance of each component is also taken into consideration; and finally, trained raters would characterize a performance by allocating a grade or rating.

Rating procedures have been repeatedly criticized for their lack of validity and reliability. First of all, as discussed in Skehan's model (Figure 3.4), performance on task appears to be directly influenced by both the rater and the rating procedures. Brindley (1994) also points out that reliability of ratings is a major problem in task-based assessment for the classroom. He argues that, a rater's subjective judgements of language performance heavily affect task-based assessment. Although rater training is the procedure usually followed to ensure reliability, Brindley (1994) reports that a rater's tendency for severity or leniency in judgements seems to remain unchanged.

Leung and Teasdale (1997) also argue that teachers-as-raters draw upon a range of different professional and personal interpretations, outside the level descriptions, to arrive at judgements about language use. They raise a question about " the degree to which it can be taken on trust that raters conceptualize construct and its attendant universe of content in the same way" (p. 68).

Furthermore, feasibility of obtaining inter-rater reliability with respect to language performance has come increasingly under question. Research in language testing has shown that despite training, significant and substantial differences between raters persist and that rater behaviour can change significantly over time (Lumely and McNamara, 1995). Although the use of the new advances in measurement technology, such as Multi-faceted Rasch analysis, has contributed to arriving at the test-takers ability independently of the raters and the rating scale, it has not been able to diminish completely the effects of such raters and rating scales (Brindley, 1994). Despite several problems that are inherent to the "rating procedures" method of measuring performance in TBA, it is still widely used both in LT research and in actual language testing situations.

3.7.4.2 Analytic Detailed Measures. This approach to measuring performance on tasks is a rather recent approach employed by a number of researchers in TBA. Analytic detailed measures provide counts of specific linguistic features occurring in test-takers' discourse as a result of performing the task. The analytic measures differ from rating procedures in that once the measures have been identified and applied to the data, there will be a considerable amount of consistency in the processing and analysis of the data.

In this research, the analytic detailed measures have been employed in some TBA research. The features that are measured usually vary in different studies in order to serve the specific purpose of each study. SLA and LT researchers have mainly employed various analytic measures to investigate accuracy, complexity, fluency, lexical density, appropriacy of language use, and indicators of negotiation of meaning in language performance on tasks.

Skehan (1998, 2001) indicates that the implications for the evaluation of performance on tasks are based on what he sees as the three primary goals of a task-based language syllabus: accuracy, complexity, and fluency in communication. Norris et al. (1998) draw upon Skehan's proposal and emphasize that, in addition to the primary goals of task-based syllabus, a principled analysis of task difficulty component is also essential for task-based assessment. Norris et al., (1998) choose the same three primary goals of accuracy, complexity and fluency as their task performance variables. However, they adopt a broader perspective into each variable by rating each of these variables in terms of the task-specific requirements for their involvement in achieving success. Based on this L2 performance perspective, they propose:

1. Accuracy would involve the minimum level of precision in code usage necessary for successful communication.
2. Complexity would involve the minimum range of grammatical/structural code required for successful completion of a given communication task.
3. Fluency would involve the minimum on-line flow required by a given task for successful (acceptable) communication. (pp. 58-9)

They state that each minimum level would necessarily be determined according to real-world criteria as identified by needs analysis and not based on native-like performance standards, as Skehan suggests. Although the framework Norris et al.

propose appears to be comprehensive, such needs analysis has not been conducted yet. Therefore, given the limited scope of the present research, it is not possible to carry on the needs analysis or to adopt the approach suggested by Norris et al. (1998). Skehan's proposed framework for task difficulty (discussed in detail in Chapter II) would offer a principled means for categorizing ability requirements and task characteristics that are inherent in L2 tasks. A number of researchers in task-based research (Brown et al., 2002; Ellis, 2003; Norris et al., 1998, 2002) support this proposal and agree that by identifying these components within a task, variable sources of difficulty will be estimated. Norris et al. (1998) suggest that with such a system for estimating task difficulty, learner performances on carefully selected tasks can be used to predict future performances on tasks that share similar task characteristics. They argue that empirical support for a system like this can lead to much improved generalizability for task-based assessment. This approach to measuring performance in TBA has not been welcomed by LT organizations, as it is time-consuming and uneconomical. Although such a framework is still at an early stage of its development and use, it has been employed by a number of researchers in TBA and is selected as the theoretical framework for task-based language assessment in the current research.

3.7.5 Reliability, Validity and Authenticity in TBA

Task-based language assessment is generally known as an advanced type of assessment with certain recognized features. It is known to be authentic and contextualized, it provides positive washback effects, it has apparent construct validity, and it is potentially suitable for generalizing the test results. However, there are certain threats to or problems with TBA.

The first problem with TBA is the reliability of the rating procedures adopted by many researchers. As discussed before, rating scales are directly influenced by different people who would use them, including scale constructors, actual raters and consumers of scores (Alderson, 1991). Another threat to reliability, as proposed by Norris et al. (1998), deals with the extent to which the real-world criteria for rating task success are based on a reliable analysis of actual judgements of task success in a real-world situation.

One of the assumptions made about TBA is that, since the communicative tasks used for assessment are based on authentic language use, they are valid. However, this assumption is criticized for a number of reasons. Brindley (1994) and Spolsky (1985) contend that in the first place, an assessment activity is by itself an artificial situation; no matter how 'life-like' the task is, people still know they are being assessed. Norris et al. (1998) refer to the same point and argue that even though an assessment task may be authentic, it does not mean that the sampling of the language is sufficient to make generalizations to other language use situations.

Brindley (1994) points to another problem with validity in TBA: how to define the criteria for assessment, upon which a test-taker's performances will be ranked or scored. Traditionally expert judgements have been used to pinpoint key elements for assessment, distinguishing different levels of student performance, and provide descriptors, ratings and so on. However, as Norris et al. (1998) argue, "experts often disagree, based on their own background and personal construct of language ability, resulting more in going round in circles than achieving consensus" (p.63). Brindley (1994) further mentions the disagreement between developers and teachers as well as among teachers themselves. As a result, he suggests that a list of assessment criteria might end up being quite arbitrary and superficial.

Another way to define assessment criteria is to utilize rating scales which already exist and are readily available (e.g. ACTFL or IELTS scales as explained earlier in this chapter). However, numerous problems exist with these scales. Lack of empirical support and difficulty in distinguishing clearly between levels, or the scales being too general to be applied to all tasks are some of the problems with such standardized scales (Alderson and Banerjee, 2003a; Norris et al., 1998).

A third problem with TBA is the difficulty of generalizing from one-off performance to other situations of language use. A crucial query for language testers to deal with is how to generalize from finite task-based performances to other potential real-world tasks. Questions such as how to determine task-inherent ability requirements, and how to evaluate learner performance within the range of such ability requirements as well as task characteristics would all influence generalizability in TBA. Bachman (2002) states that " [I]nconsistencies across tasks affect generalizability, or the extent to which our inferences generalize across a set of assessment tasks" (p. 458).

Finally, there are a number of practical issues that TBA has to take into account. As mentioned before, TBA is time consuming and uneconomical since it involves eliciting, evaluating and scoring student performances one at a time.

As the discussion in the previous sections demonstrates, it is clear that use of tasks and TBA can be, to a large extent, justified, in terms of validity and authenticity within the communicative framework of language testing. However, there are certain problems with TBA which are to be addressed in this research. As regards generalizability of the test results, it is hoped that with careful estimations of task difficulty, language testing can help identify a clear framework in which learner performances on carefully selected tasks can be more accurately measured. In effect, one purpose of the present research is, through investigations of different task

characteristics, to provide a clearer index of task difficulty. Determining task difficulty would inevitably provide LT researchers with a better indication for generalizing results of an assessment situation to similar language use situations. As discussed before, another problem in TBA is the reliability of the rating scales and rating systems that are usually used in evaluating oral performance. To achieve more reliability with the test scores in the present research, therefore, the use of rating scales is avoided. Instead, detailed analytic measures are adopted for assessing the test-takers' performance on task, which will be thoroughly discussed in Chapter V.

CHAPTER IV

Variables in Task-Based Research

4.1 Introduction

During the last two decades, tasks have become a popular means of language teaching and testing. As discussed in Chapter II, research in task-based language teaching (Loschky & Bley-Vorman, 1993; Ellis, 2003) has demonstrated that tasks are useful devices to practise language as a tool of communication rather than as a device to focus on grammatical features of the language. Skehan (1998) proposes that the focus on meaning in tasks "will engage naturalistic acquisitional mechanisms, cause the underlying interlanguage system be stretched, and drive development forward" (p. 95). Research into task-based instruction has indicated that different characteristics of tasks can influence task difficulty, which in turn might have an intensive influence on language performance in terms of its quantity and quality. The cognitive load of a task, its communicative stress and its linguistic complexity are reported as three significant factors affecting the difficulty level of the task (Skehan, 1996, 1998). Investigations have shown that familiarity with task type and content, chronological sequence, availability of pre-task planning, and the learner's level of language proficiency are all factors that would affect the difficulty associated with performance on tasks. In the present chapter, I will focus mainly on how task characteristics and performance conditions influence task performance. I will explore task structure, pre-task planning time, level of language proficiency and learner perceptions of task

difficulty in some detail because these are generally regarded as important factors in task performance.

4.2 Task Structure

A number of recent studies have looked at task structure as a significant characteristic of oral narrative tasks and its impact on language performance in a teaching or testing setting (Skehan and Foster, 1998; Wigglesworth, 2001). Although structure of a task has been introduced and defined in these studies, not much attention has been paid to establishing a systematic definition for task structure. Some of the studies have not clearly problematized what 'structure' in the task-based context refers to or how it is defined and operationalized. In addition, little consensus can be seen among interpretations of task structure in task-based studies. As a result, a fundamental contribution of the present study is to consider task structure in the broader context of SLA literature and subsequently present a more systematic description, interpretation and operationalization of task structure. To achieve this purpose, first a summary of the recent literature on the concept of "structure" is given. Drawing on SLA literature, I will then focus on how task structure is defined and operationalized in the present study.

4.2.1 Structure in Task-Based Research

Foster and Skehan (1996) and Skehan and Foster (1997) in their studies on the effect of task characteristics on learner performance have explained the idea of task structure in a general sense. In these two studies, the inherent structure of a task is defined in terms of time sequencing and degree of organization of input material. They argue that presence of structure in a task would help to ease the processing burden of a task

on learners. For instance, in Foster and Skehan (1996), the "oven" task -a personal instruction-giving task in which the narrator had to instruct a friend to get to his/her home, get into the kitchen and turn the oven off - is considered to be a structured task since the instructions to be given occur in a non-arbitrary sequence and the information is familiar to the person who is performing the task. The structure of the task, in this case, is manifested through both the sequencing of the task, i.e. each step in the sequence leads to the next in a familiar way to the narrator, and the familiarity of the input information. Structure, in Foster and Skehan's study, is introduced as a task characteristic that influences learner performance. However, since it is associated with other task characteristics such as familiarity of information or the direction giving nature of the task, they did not operationalized structure as an independent characteristic of a task.

Operationalizing task structure in a rather different way, Skehan and Foster (1997) have used a "Sempe" cartoon strip task, a narrative task based on picture stories, and asked the learners to look at the pictures and tell the story to their partners. In this study, structure of the oral narrative task is demonstrated through the clear macrostructure and the amusing punch line of the story, but in this case the information is less familiar to the learners. Skehan and Foster discuss that clear sequential structuring of a task, which is realized through time sequencing, represents the macrostructure. Based on their previous studies, Skehan and Foster (1999) have attempted to provide a more detailed definition of structure. They define the structure of a narrative task in terms of the schematic knowledge of the interconnected events and the predictability of the sequence of the actions that happen in the narrative. In Skehan and Foster's (1999) study, participants were required to retell narratives based on two video prompts: Mr Bean episodes of 'Restaurant' and 'Crazy Golf'. The

'Restaurant' episode represents a relatively structured narrative in which all the events happen in a clear and predictable sequence, i.e. "the sequence of the actions is predictable and follows a fairly necessary path" (Skehan and Foster, 1999, p. 104). In addition, this narrative is considered structured since it provides the participants with a familiar "restaurant script[1]". The "Crazy Golf" episode, in contrast, is a relatively unstructured task in which the sequence of events is unpredictable and the events are not interconnected. Skehan (2001), in a general overview, further develops the concept of task structure and considers it as clear macrostructure, with the time line sequence underlying the task. This time line for the information underlying the task is clear and of significance to the overall content organization of the task. Therefore, it appears that the macrostructure of the task refers to the schematic knowledge the learners have about the information that is provided in the task.

Foster and Skehan (1996) and Skehan and Foster (1997, 1999) have considered structure as a representation of macrostructure with specific attention to time sequential organization in the case of narrative tasks. They have looked at structure from a reasonably systematic view and provided logical interpretations of clear macrostructure. However, it should be mentioned that the way they have considered and operationalized structure is not the only way of defining task structure in the context of SLA in general and task-based research in particular.

For example, Wigglesworth (2001) views structure as a task characteristic that influences the cognitive complexity of a task. In her study, she defines structure as the amount of information provided to test-takers to assist them in performing the task. She operationalizes structure in terms of the number of specific prompts given

---

[1]The information we possess as a background to what we comprehend has been called a *script.* In a 'restaurant script', for example we know that at a restaurant, one sits down at a table; waiter brings the menu; one orders; waiter goes away to order the main dish and so forth.

to the test-takers to direct them in their interaction with their interlocutors. In fact, in Wigglesworth's (2001) study, a task is structured if the participants are provided with five specific prompts on how to engage in their interaction with their interlocutors. In contrast, a task is considered unstructured if only one general statement is given to guide the participants in a task. Although Wigglesworth quotes from Candlin (1987) that the chronological sequencing of a task is a contributing factor to the cognitive complexity of a task, she does not consider this definition as the underlying concept of 'structure' in her study. It should be noted that there are some uncertain points in the way Wigglesworth (2001) defines and thus operationalizes task structure. First, in her definitions of task difficulty she builds on the assumptions made by Candlin (1987) and Skehan (1998), but she adopts neither assumption. Identifying variables of the study, Wigglesworth (2001) explains that:

> The task was developed either with or without structure. This was operationalized in terms of the amount of information provided to the learners to assist them in doing the task. Specifically where structure was present the learners were provided with five specific prompts to direct them in their interaction with their interlocutor. Where structure was not provided, one general statement was provided to guide the learners in the task (p. 191).

This explanation does not clearly define the relationship between structure and information. Nor does it elucidate when structure is disturbed through lack of adequate information, what type of and how much information is missing. The definition she provides for task structure seems to be very similar to the concept of "adequacy" raised by other researchers (Iwashita et al., 2001). However, she does not use the term 'adequacy' (and it is unfortunate that this discrepancy in the literature remains unresolved). Finally, she operationalizes structure by the number of

statements provided to the participants to direct them in their interaction with their interlocutors but this operationalization of structure is neither explained nor justified. In other words, it is not clear what makes a task unstructured, which prompts are excluded in the unstructured version of the tasks, or which prompts are kept in the structured task.

Iwashita et al. (2001) have manipulated the concept of task difficulty through the four dimensions of perspective, adequacy, immediacy and planning in an assessment setting. They have used oral narrative tasks based on a sequenced set of pictures as the stimulus routinely used in the Test of Spoken English (TSE) to elicit test-takers' language performance. In their study, perspective was defined as the way in which the participants told the story from their own point of view or from someone else's point of view. Immediacy, following Rahimpour (1997) and Robinson (1995), was described in terms of the presence or absence of the picture sets in front of the test-takers when they were telling the story. Planning time was manipulated by providing the test-takers with either 0.5 minutes under unplanned conditions and 3.5 minutes under planned conditions. The last task dimension was adequacy which referred to the conditions of telling the story with a complete set of pictures or with an incomplete set of pictures. In fact, adequacy in this context seems to refer to the amount of information required to make a task more or less difficult to complete. This characteristic of oral narrative tasks appears to be what Wigglesworth (2001) has called 'task structure'. Therefore, structured tasks in Iwashita et al. (2001) appear to be those in which test-takers narrate the story with a complete set of six pictures, whereas the unstructured tasks are narrated with an incomplete set of four pictures.

Like Wigglesworth (2001), Iwashita et al. (2001) have neither clearly defined the concept of adequacy nor identified the rationale for employing or operationalizing

adequacy in their study. Iwashita et al. (2001) did not discuss whether their 'adequacy' refers to Wigglesworth's (2001) concept of structure or to Robinson's (1995) and Rahimpour's (1997) concept of here-and-now versus there-and then. The questions of what type of and how much information is essential for a task to be considered structured also remain unanswered. Iwashita et al. (2001) have not dealt with the underlying principles of structure, in this sense, and have not adopted a systematic approach to including or excluding a number of pictures in each picture set. The only explanation they have provided is that in the 'inadequate' tasks two of the pictures are missing. However, they did not mention which two pictures are missing or how the pictures fit in and relate to the other pictures in a series.

Investigations of the concept of 'task structure' in task-based studies show that a good deal of research is required to explore the underlying principles of structure in the broader context of SLA. Defining task structure and the way it is operationalized essentially seems to be a challenge to task-based research. Therefore, a fundamental purpose of the research reported here is to explore recurrent concepts of structure in first and second language acquisition on the basis of which this study can be established. As regards the purpose of this study, the structure of oral narrative tasks[2] should be carefully considered, examined and defined. In the following section two recurrent approaches to the concept of structure within the context of SLA will be explained and discussed.

4.2.2 Structure in Language Acquisition Literature

In the literature on task-based language teaching and testing, a number of research

---

[2]In the current discussions of TBA and SLA and in the present study, structure of an oral narrative task generally refers to the structure of the picture stories and not the task itself.

studies have attempted to explore what structure is, how it influences language learning and how it can be manipulated. In the following section, I will focus on two recurrent approaches to contextualizing and defining structure.

4.2.2.1 Structure: A problem-Solution Concept. Winter (1976), working on different information structures of English discourses, provided a framework for information text structure within the context of teaching students how to communicate efficiently through their written work. In an attempt to analyze the concept of information structure, he investigated a number of different texts to find out how structure is characteristically represented in those texts. In his initial discussion of the communication between writers and readers of a text, he points out the fact that the relation between what the readers do not know and what they do is the bridge which enables communication to take place. He states that within a text, readers are most likely to ask for an explanation of something which they do not understand or they need to know for additional related information. Drawing on the concept of communication in a larger context, he claims that communicators normally ask four questions which dominate all others: " What is the situation?; What is the problem to be solved?; What is the solution?; and How is the solution to be evaluated?" (Winter, 1976, p. 2). Such information structure, he goes on to say, might present itself in three patterns.

Pattern 1:	Situation	Problem	Solution	Evaluation
Pattern 2:	Situation	Problem		Evaluation
Pattern 3:	Situation			Evaluation

He states that in Pattern 2, a likely solution could be proposed in the evaluation and in Pattern 3 a likely problem could be raised through evaluation. He further discusses

that this type of structure, i.e. problem-solution structure, represents information in its most general and fundamental way and it is the recurrent pattern in which all technical articles and reports are presented. Pointing out the significance of problem-solution structure to readers and writers, Winter (1976) argues that:

> An important part of communication in Science and Technology is about Problem and Solution, and within these two it is primarily concerned with notions of efficiency in HOW and WHY things are done. Students are quite happy with nice solid information about Problem and Solution but such information does not in itself constitute adequate communication for the innocent reader or listener. It is the complementary comprehension functions of Situation and Evaluation that ensure the reader or listener will understand the significance of Problem or Problem-Solution. (p. 16)

Hoey (1983), developing ideas formulated by Winter (1976), discusses different forms of information structure and the common assumptions of discourses. He states that discourses and passages of discourses are organised, or at least organised in part, in a hierarchical manner. He argues that native-speakers of a language can assess whether a sentence is grammatical as equally as they could recognize whether a discourse is well formed. He contends that there is something in the discourse that helps the listener or speaker perceive its structure. Hoey (1983) argues that the problem-solution pattern is a common discourse pattern in English which, in its full form, may be demonstrated in the sequence of: " situation, problem, response, result, and evaluation". He defines 'situation' as the setting in which events happen and the problem occurs. 'Response', in his definitions, refers to the solution being made for the problem that, in turn, will produce a result. And 'evaluation' reflects the resolutions being made on the results emanated from the response. A positive

evaluation of a response will satisfactorily round off a discourse, whereas a negative evaluation would signal another problem, and in fact, initiate a new problem-solution sequence. Both Winter (1976) and Hoey (1983) propose that in order to be easily understood, information texts would ideally have problem-solution structures. The following extract shows an example of a problem-solution structure provided by Winter (1976):

The Typical Anecdote Structure

Situation?	I was walking along a country road some time ago.
Problem?	A lorry mounted the pavement and came straight for me.
Solution?	I threw myself in a hedge to get out of its way.
Evaluation?	The lorry missed me by inches.
	What a lucky escape I had. (p. 8)

Winter (1976) called this an anecdotal use of problem-solution structure but emphasized that problem-solution structure is also recurrently used in science and technical discourses as well. Interestingly, the example given by Winter is a problem-solution structure within a narrative which resembles the typical oral narrative tasks employed in task-based studies.

Besides Winter and Hoey, there are other researchers who have introduced the problem-solution sequence as a significant pattern of text structure. Richgels, McGee, Lomax and Sheard (1987), in their study on the effect of text structure on recall, have introduced four main categories of text structure: collection, comparison-contrast, causation and problem-solution. They have reported that learner awareness of such text structures would facilitate recall of texts written in those structures. In the case of narratives, Richgels et al. (1987) reported that learners benefited from problem-solution structure and were able to recall the written texts of problem-solution

structure more successfully. They discuss that presence of problem-solution structure in a narrative would strengthen learners' ability to recall when they are retelling the texts. Turner (1992), drawing on cognitive psychology, defines problem-solution as one of the significant 'top-level structures' of thinking and considers it as an integral part of organizing and dealing with aspects of everyday life. She claims that these structures help us to link present and past experiences, make decisions, solve problems, enrich and expand our understanding of concept and appreciate critically aspects of our world.

Investigating structure in a language testing context, Kobayashi (1995, 2002) showed that problem-solution structured reading comprehension texts produced different results to those which were not structured in this way in that they distinguished more clearly between levels of performance based on the reading passages. Higher proficiency students, in other words, were more able to respond to the structure within the texts, especially when given comprehension tasks of a more demanding nature. In effect, presence of a problem-solution structure in the reading comprehension texts helped higher proficiency learners perform more successfully in their tests.

As the discussions presented in this section indicate, SLA researchers have investigated 'structure' within the context of texts. This body of research suggests that a frequent structure pattern in narratives, as reported by a number of researchers, is the problem-solution structure. The results of these studies clearly indicate that the presence of a problem-solution relationship in a narrative would help language learners to comprehend, recall and perform on the text more efficiently. It also suggests that information structure in general and problem-solution structure in particular would organize texts and would accordingly have a facilitating role on learners' comprehension and performance. It appears that learners, having observed a

problem, would start hypothesizing the possible solutions and evaluations which would actively engage them in the situation. I would also argue that problem-solution structure has a facilitative role in comprehension because it is part of a human's schemata. In fact, our knowledge of world and experience of life is, to a great extent, based on a problem-solution pattern. As a result, it could be hypothesized that a problem-solution structure of an oral narrative task might further have facilitative effect on language performance on the tasks.

4.2.2.2 Structure: A Schematic Concept. Mohan (1991) has looked at the nature and types of knowledge structures in the context of second language for academic purposes. Drawing upon the research done in the field of cognitive psychology (Schank & Ableson, 1977), anthropology (Malinowski, 1935) and genre linguistics (Martin, 1985; Swales, 1985), Mohan expands on the idea of 'knowledge structure'. Before dealing with his discussions and definitions of structure, it should be noted that Mohan presumes knowledge structures as text structures or genres (p. 15). Discussing the importance of 'knowledge structures' and 'student tasks', Mohan (1991) argues that:

> Research on 'knowledge structure,' or information patterns, provides evidence that they are cross-cultural, that they underlie subject-area knowledge and thinking skills, that they can be represented by graphics, that they underlie expository reading and writing, being realized in discourse and grammar in a variety of ways, and that student awareness of them improves retention of subject matters". (p. 2)

Therefore, knowledge structures, as Mohan describes, are patterns of organization and are important in both language and content knowledge areas. In his research in the

field of language pedagogy, he attempts to investigate how learners organize knowledge to understand, remember and apply new information. From a cognitive psychology point of view, he argues that knowledge, including language knowledge, is schematized or organized in chunks or packages. The existence of such organized schemata or knowledge structures would then facilitate comprehension, memory and application in both reading and writing skills. He further contends that knowledge structures are cross-cultural and therefore appropriate to learners from a range of different cultural backgrounds.

Mohan (1991) categorizes knowledge structures in three pairs of related patterns: (1) a description of a particular object or person that involves classification or set of general concepts; (2) a particular temporal sequence of states, events or actions that often involve general principles and relate one state to others; and (3) a particular choice or decision that often involves general values. Mohan (2001) modifies this classification into six core knowledge structures of description, sequence, classification, principles, values and choices. He claims that all English discourses would ideally be placed in one of these six knowledge structures.

Mohan (1991, 2001) has proposed this framework to aid the development of communication, thinking and language, and to facilitate teaching of themes and topics to second language learners. However, as regards the present study, his definition and classification of narratives in knowledge structures is important. Mohan (1991) considers narratives, action strips and time lines as examples of graphic representations of temporal sequence and places them in his classification of "temporal sequence of events and action". Narratives, in this framework, are considered as a basic type of knowledge structure, are schematized and organized on the basis of social rules or cause-effect relations, and would aid the development of

language and communication. In Mohan's description of knowledge structure, therefore, the schematic structure of a narrative and the logical time sequencing underlying the events are the two major conditions for structured narratives. The schematic structure, in this sense, is reflected through the general principles of schema including the content schema and a formal schema. Content schema, as explained by Mohan (1991), refers to knowledge relative to the content domain of the text and formal schema refers knowledge relative to the formal organization of structures of different text types. In addition, the time sequencing is demonstrated by the temporal sequence of events, states or actions which relate one state to the others.

Another research area which has dealt with the effect of text structure, e.g. narratives, on language learning is the studies carried out on learner recall in their first language reading skill. Mandler (1978) in a study on the role of some characteristics of schemata in encoding and retrieval of stories investigated the effect of structure on learner recall. She generated two-episode stories in two different versions whose underlying narrative structure was violated by "interleaving" the events of the two episodes. While studying the recall of the learners through their production of the stories, she found that quantity of the recall for the interleaved stories was far less and its quality was linguistically marked. The results of her study revealed that the recall of the two versions were different from one another with the most pronounced effect being found due to sequencing. In other words, the lack of a proper schematic sequencing was found to be the main reason for having limited amount of linguistically marked recall. Mandler (1978) argued that the structure of a narrative, particularly its sequencing, influenced comprehension and recall since the operations of schemata were directly depended on such structures. Mandler (1978) concluded that narrative schemata consist of sets of expectations about stories, the units of which

they are composed, the way in which those units are sequenced and the types of connections between units that are likely to occur. She states that disturbing any of these expectations in a narrative would have certain negative effects on comprehension and recall of the text.

Carrell (1985), Meyer, Brandt, and Bluth (1980) and Kintch and Van Dijk (1978) have all worked on the effect of schematic structure on second language reading comprehension and/or recall. The results of their studies reveal the fact that schematic structure of a text plays an important role in discourse comprehension and production. Carrell (1985) defines structure as the rhetorical organization of a text which interacts with the learner's schemata, background knowledge and experience. In a narrative text, she states that a hierarchical schematic structure is present to which both native and non-native readers are sensitive. Similarly, Meyer et al. (1980) claim that the structure has the function of specifying the logical connections among ideas or events in a text as well as the subordination of some ideas to others. In the case of a narrative, they all mention sequencing as a salient enabling organizational factor.

To summarize, the above discussion suggests that two of the most frequently reported approaches to defining the notion of structure of a narrative in SLA literature are schematic sequential organization and problem-solution (These two approaches to defining structure do not exhaust the different ways in which the structure of an oral narrative task can be defined.). A schematic sequential structure is one in which different events or states, following a clear timeline, occur in an organized temporal sequence, where each event or state is based on the one that comes before and is essential for the events or states that follow. A problem-solution structure is a type of structure that, in addition to the timeline and sequential organization, presents a problem-solution relation to the reader of the text. In effect, in a problem-solution

structure there is a situation in which a problem is exposed, a solution is suggested to the problem (usually by the participants of the story) and eventually an evaluation is made based on the outcome. This sequence of situation, problem, solution and evaluation is itself a reflection of the time line and sequential organization of the narrative. Therefore, it is clear that a problem-solution structure would inevitably contain, at least to some extent, the schematic sequential structure as well.

It is worth mentioning that, for an oral narrative task (picture stories), being structured or unstructured is not a matter of a dichotomy. Instead, structure of a narrative should be viewed as a characteristic that spans along a continuum. In other words, it is more sensible to compare different types of narrative tasks and find varying degrees of structure that can be placed along a continuum which ranges from unstructured to structured. The two types of structure introduced in this study, therefore, could be distinguished from one another for the degree of structure they expose. For instance, comparing a problem-solution with a schematic structure, it could be proposed that a narrative that is based on a problem-solution structure would have a higher degree of structure than a narrative which is based on a schematic sequential structure. The reason, one could argue, is that a problem-solution narrative exposes both a problem-solution relation and, at least to a reasonable extent, a fixed sequential organization of events (as discussed in the previous paragraph).

As the discussion presented in this section suggests, both problem-solution and schematic sequential structure influence second language learner comprehension, recall and performance. Although a number of researchers have looked at structure from a text or genre perspective, rather than a task approach, structure appears to denote the same concept in both perspectives and to be based on the same underlying principles. The significant conclusion to be drawn from the studies reported here is

that the presence of structure promotes learner comprehension and performance. Another crucial point these studies make is that disturbing structure would have a debilitating effect on second language performance. As regards the purpose of the present study, therefore, it can be hypothesized that the structure of an oral narrative task would influence language performance on tasks. Furthermore, as suggested by the discussions in the previous paragraphs, it could be hypothesized that a problem-solution structure exposes a higher degree of structure than a schematic sequential structure on tasks and on language performance.

In the section that follows, I will explain how structure is operationalized in the present study and how different tasks are selected to represent varying degree of structure. The detailed description of each individual task employed in Studies One and Two will be presented in Chapters V and VIII respectively.

4.2.2.3 Operationalizing Structure in the Present Study. Based on the discussions of structure presented in the previous section, two types of structure, i.e. problem-solution and schematic sequential, were employed. In operationalization of the structure of oral narrative tasks in this study, the presence of a problem-solution structure and/or a schematic sequential organization would indicate that a task is structured. Conversely, the lack of a problem-solution relation or a schematic sequential organization with a clear time line in an oral narrative task would suggest that the task is unstructured (or less structured[3]). To have an opportunity to compare varying degrees of task structure and drawing upon the above-mentioned approaches to defining structure, four types of tasks with varying degrees of structure are defined.

---

[3]For writing purposes, the term 'unstructured tasks' will be used in the current discussion to refer to tasks which are less structured.

Figure 4.1 shows how tasks of different degrees of structure are placed along a continuum of structure in the present study.

Figure 4.1

Degrees of Task Structure

+ Structured		**- Structured**	
1. Problem-solution	2. Schematic	3. Loose Sequential Organization	4. Unclear Time Line

As Figure 4.1 demonstrates, two of the tasks are considered as structured (Tasks 1 and 2) and the other two are assumed to be unstructured (Tasks 3 and 4). The type of structure the structured tasks expose discerns them from one another. As discussed in the previous section, a problem-solution task is taken to be more structured than a schematic sequential task since it contains both a problem-solution relation and an organized sequence of events (i.e. situation, problem, solution and evaluation).

With the unstructured (or less structured) group, neither a problem-solution relation nor a schematic sequential organization is exposed by the narrative tasks. Tasks that are considered unstructured have either a loosely presented sequential organization or an unclear time line (which, in fact, means no sequential organization) with an arbitrary relationship among different events and states in the narrative. Since a clear time line is not exposed in the unstructured tasks, each event or state is not dependent on the one that comes before or after it and, therefore, could happen in different sequence. As a result, in an unstructured task, there are events that could be rearranged without the main theme of the story being changed. In effect, one significant facet of operationalizing structure in this study is the number of pictures that can be rearranged in a narrative without the main story being compromised. In the case of a structured task, each picture has a key role in sequencing and the

ordering of the pictures is non-negotiable. If, for instance, one picture is taken out or replaced by another one the main theme of the story would change. In contrast, in an unstructured task, one or some of the pictures could be rearranged without any major change in the main theme or outcome of the story. The detailed discussions about each individual narrative which is selected to represent different types and degreed of structure will be presented in Chapters V and VIII.

4.3 Pre-Task Planning in Task-Based Research

As indicated earlier in section 4.1, pre-task planning is reported to have an impact upon task difficulty and performance on tasks in task-based studies (Mehnert, 1998; Ortega, 1995; Yuan & Ellis, 2003). Pre-task planning is considered as a performance condition in task-based studies and is known to have a direct influence on different aspects of L2 performance. In the past two decades, a number of SLA studies have looked at pre-task planning time and investigated its detailed effects on second language performance. In general, the results have shown that giving learners and test-takers some planning time before they perform the tasks has a favorable effect on their performance. In the sections that follow, I will explain how planning time is operationalized in different task-based studies and what effects pre-task planning would have on L2 performance. I will finally present a table which shows the results of studies investigating pre-task planning in language teaching and testing contexts.

4.3.1 Operationalizing Pre-Task Planning

Regardless of the specific purpose of the studies in language teaching or testing, speech planning has usually been operationalized in terms of the amount of time given to a learner or test-taker in advance of task performance. Studies vary in terms of the

amount of planning time and the focus or method of implementing pre-task planning. Focusing on the effect of planning, most SLA studies have worked on adult learners performing one or more tasks in a classroom setting. Different types of tasks including narrative, picture description, decision-making, personal information exchange and instruction-giving are used. It is worth mentioning that in task-based instructional settings, researchers have employed a wide range of tasks to suit the purpose of their studies. However, in assessment settings, because issues such as authenticity of tasks and practicality of tests are to be taken into account, narratives have become more popular than other task types.

The operationalization of planning time in studies that have been carried out in instructional settings ranges from one to ten minutes, with a number of studies incorporating a ten-minute planning time condition. However, it appears that the amount of planning time in SLA studies is strongly governed by the purpose of each individual research study. Researchers who are investigating the effect of planning time on language performance in a teaching situation have usually selected a longer planning time. Foster and Skehan, (1996) and Skehan and Foster (1997), following Crookes (1989), have provided the learners with 10 minutes of planning time. Ortega (1999) has also operationalized planning in a 10-minute condition. Mehnert (1998) has compared the effects of 1-minute, 5-minute, and 10-minute planning time in a teaching setting. In contrast, researchers investigating planning time from an assessment perspective, considering test authenticity and validity as well as practical restrictions of testing, have employed shorter amounts of planning time in their studies. Wigglesworth (1997) provided the candidates with a 1-minute planning time before taking the test. Iwashita et al. (2001) and Elder et al. (2002) have operationalized planning in a 3-minute condition. Although a 3-minute planning time

is practical from a testing point of view, results of some recent studies suggest that more investigations are needed to find the effect of varying amount of planning time on language performance in assessment settings (Elder et al., 2002; Iwashita et al., 2001; Wigglesworth, 2001).

The method of operationalizing planning time also varies across various studies. Ellis (1987) operationalized planning as an opportunity to monitor and plan during writing with a further chance to rehearse or to become familiar with a task when telling a story orally after having written it once. Following Ellis, Ting (1996) used planning time for a bimodal task of writing versus speaking. All subsequent studies of planning time have concentrated on tasks of speaking and investigated the effect of pre-task planning on oral performance. Foster and Skehan (1996) have compared different operationalization methods of planning in a "detailed" versus "undetailed" planning condition. Under the detailed planning conditions, the learners were given some suggestions on how to use planning time to consider lexis, syntax and so on. Under the undetailed planning conditions, on the other hand, the learners were simply told to plan. Foster and Skehan (1999) have operationalized planning in terms of the source and the focus of planning. Concerning the source of planning, three different groups of learners were formed based on whether they belonged to the group-based, teacher fronted or solitary planning conditions. Each of the groups was also directed to focus on either planning for language or planning for content. In other studies, however, the learners are given the planning time and just asked to plan.

As the discussions in the previous paragraphs suggest, operationalization of pre-task planning depends, to a great extent, on the pedagogic or administration settings of the study. Taking an assessment perspective, this study will attempt to investigate the effect of planning time on second language performance in line with the practical

restrictions of assessment settings. The amount and method of implementing planning time would, therefore, be adopted with reference to the considerations of practical timing issues. As discussed earlier, the method of implementing planning should also be appropriate for assessment purposes. The results of Mehnert's (1998) study revealed that a larger amount of planning time could, to a great extent, enhance different aspects of performance. The 1-minute planners were only different from the 5-minutes and 10-minutes planners in terms of the accuracy of their performance. However, with more planning time available to the learners they were able to improve different aspects of fluency and complexity of the language they produced. A number of studies performed in assessment settings reported little or no effect of a 3-minute planning time on L2 performance (Iwashita et al., 2001; Elder et al., 2002).

As the above discussion indicates, LT researchers tend to give a shorter planning time (3 minutes) while researchers in language teaching contexts appear to favour longer planning times (10 minutes). Under the circumstances in which the data for this study were collected, it was not possible to adopt a 10-minutes planning time. Moreover, it was important to take account of the typical planning time given to the learners in other research projects in order for the findings to be more or less relevant to those of the current study. Hence, a 5-minutes planning time was selected. Regarding the implementation of planning time, since the current study is carried out in an assessment rather than an instructional setting, the pre-task planning would be given to the learners in an undetailed and unguided method.

4.3.2 Effects of Pre-Task Planning on Language Performance

Investigating the influence of planning on second language performance, Foster and Skehan (1996) used three different tasks –a personal information exchange, a narrative task and a decision-making task. The results of their study provided

evidence for an interaction between planning conditions and task type. Under planned conditions, fluency and complexity were greater for the more cognitively demanding tasks. Surprisingly, with accuracy the greater effects were achieved in an unplanned condition. These results suggested that greater accuracy was achieved on tasks that required the least cognitive effort, and therefore, allowed more attention to be devoted to form. In this study, Foster and Skehan concluded that planning does not operate in the same way with all tasks and called for more research to be carried out to account for different effects planning time would have on performance. Similar findings were reported by Skehan and Foster (1997) who investigated the effect of planning time and inherent structure of tasks on language performance. The results of their study revealed that planning had greater effects on accuracy in tasks which contain a clear inherent structure (e.g. narratives that are based on a sequenced cartoon strip). Furthermore, planning was associated with greater complexity in tasks which required more on-line processing or had complex outcomes (e.g. giving advice).

Ortega (1999), in a study on the effect of planning time, attempted to investigate whether planning opportunity results in an increased focus on form in the context of task-based, meaning-driven communication. The results of her study showed that planning time would help learners to produce more fluent and more complex language. However, no effects were identified for the lexical range. In her study, the pattern of findings for accuracy was also inconclusive. Yuan and Ellis (2003) have investigated the effect of both pre-task and on-line planning on L2 learners' performance on narrative tasks. The results show that pre-task planning enhances complexity, while on-line planning improves both accuracy and complexity.

Considering task-based research from an assessment perspective, Wigglesworth (1997) found a selective effect of planning on task type. The findings of her study

indicated that high proficiency candidates benefited from planning more when performing more difficult tasks. Wigglesworth (1997) argued that, when the cognitive load of a task is high, planning time could play a significant role in reducing this load and would help the candidates to have a more successful performance. In contrast, planning time may not be as beneficial with less cognitively demanding tasks.

Foster and Skehan (1999), in their study investigating the effect of source and focus of planning on learners' language, reported that teacher-fronted planning would result in more accurate language and solitary planning would produce the most complex language. In contrast, using planning instructions to direct attention to language or content did not lead to any significant differences in the learners' performances.

Elder et al. (2002), following Iwashita et al. (2001), have studied the effect of planning time, as an aspect of task difficulty, on test-takers' performance on oral narrative tasks in an assessment setting. Operationalizing planning time in a 3-minute condition, they have reported that no systematic variation was associated with planning conditions. However, Elder et al. (2002) have contended that lack of consistent results in their study, compared with what has been repeatedly mentioned in SLA literature, may have emerged from a number of factors. One possibility, as they mention, might be the differences between testing and pedagogic situation. They argue that another possibility might be that "the conditions of the experiment itself were not conducive to producing marked differences in the quality of the candidate performance" (p. 362). This might also be true with the amount of planning time they have given to test-takers before performing each of the tasks. It might appear that a 3-minute planning time, compared with a norm of 10-minute planning time in a teaching situation, is not enough for the test-takers to focus on different aspects of form and meaning.

By looking at task-based studies which have focused on pre-task planning time, it can be realized that planning time would have consistent and appreciable increases in the complexity and fluency of second language performance. The results also suggest that accuracy of performance is increased, for particular tasks and/or under specific conditions, when planning time is available prior to the performance. However, the results for accuracy are usually less consistent and with smaller effects. Furthermore, the results of different studies imply that there might be an interaction between pre-task planning, aspects of performance and task type. Nonetheless, the intricate nature of planning as a cognitive process and the interactions planning might have with different types of tasks would extend the need for more research to be carried out in this area. A summary of the findings of different SLA studies which have focused on the effects of pre-task planning on L2 language performance is presented in Table 4.1.

4.4 Language Proficiency in Task-Based Research

Studies of task-based language assessment are not, in principle, restricted to certain language learners at particular proficiency levels. Language learners at different proficiency levels, have to some degree, a choice about what they say and what structures they use to perform on a task. As they become advanced in their level of language proficiency they not only learn how to use L2 more efficiently but they develop strategies in how to utilize their background knowledge in performing L2 more successfully. Hence, another factor that can influence task performance is the learner level of language proficiency not only because of the direct effect of language proficiency on L2 performance but also because of the interactions language proficiency may have with task characteristics and performance conditions. In addition, it is not certain whether learners of varying proficiencies would benefit the

Table 4.1
Effects of Pre-Task Planning on Language Performance in SLA Studies

Study	Operationalization of Planning	Task type	Proficiency Level	Measuring Procedures	Design / Statistical Analysis	Effects of Planning on Language Performance
Ellis (1987)	Write/ tell/ write and tell	Narrative: (Picture stories)	Post-beginner	Analytic Measures	Repeated measures Chi-square	More accuracy under planned conditions
Crookes (1989)	10 minutes Tell/ Plan and tell	Information gap: Map/ Lego tasks	TOEFL 430-650	Analytic Measures	Repeated measures MANOVA	Greater fluency and complexity under planned conditions
Ortega (1995)	8 minutes Tell/ Plan and tell	Narrative: (Picture stories and aural stimuli)	Intermediate	Analytic Measures	Repeated measures ANOVAs	More complex and fluent language under planned conditions
Foster & Skehan (1996)	10 minutes Tell/ detailed plan and tell/ undetailed plan and tell	Personal inf. Exchange Narrative Decision making	Pre-intermediate	Analytic Measures	Between groups ANOVAs	Greater fluency, complexity and accuracy under planned conditions Detailed planning produce more complexity and fluency than undetailed planning
Skehan & Foster (1997)	Tell/ plan and tell	Personal inf. Exchange Narrative Decision making	Pre-intermediate	Analytic Measures	Between groups ANOVAs	Greater fluency, complexity and accuracy under planned conditions
Wigglesworth (1997)	1-2 minutes Pre-task planning	Different tasks on tape-mediated oral test	Low and high proficiency groups	Analytic Measures	Between groups Chi-square	More accuracy for 1-minute planning of high-intermediate planners on the more difficult task
Mehnert (1998)	1, 5, and 10 minutes Tell/ plan and tell	Phone messages: Instruction task Exposition task	Early intermediate	Analytic Measures	Between groups ANOVAs	Fluency, accuracy and complexity all improved but varied with planning conds.

Ortega (1999)	10 minutes Listen-look and tell Listen-look plan and tell	Narrative (taped version of the story followed by picture strips)	Advanced	Analytic Measures	Within groups ANOVAs	More fluency and complexity under planned conditions
Foster & Skehan (1999)	10 minutes Teacher-led/ group-based/ solitary planning Content/ language planning	Decision making	Intermediate	Analytic Measures	Between groups ANOVAs	More complexity and fluency in solitary planning More accuracy in teacher-led planning More fluency with content focus planning
Robinson (2001)	Amount of planning not specified Tell/ plan and tell	Information gap: Map task	Not defined	Analytic Measures	Repeated measures MANOVA	Planning condition, along with other task characteristics result in more lexical variety and fluency of the language
Wigglesworth (2001)	5 minutes Tell/ plan and tell	Giving instructions Negotiating an oral transact./complex spoken exchange Obtaining inf.	Two levels (not defined)	Rating scales	Between groups ANOVAs	Planning has no effect on familiar and structured tasks More complex language may be associated with planning
Iwashita, McNamara & Elder (2001)	3 minutes Tell/ plan and tell	Narrative (picture stories)	TOEFL 427-670	Rating scales & Analytical	Repeated measures ANOVAs Rasch analysis	No effects were found
Elder, Iwashita & McNamara (2002)	3 minute Tell/ plan and tell	Narrative (picture stories)	TOEFL 427-670	Rating scales	Rasch analysis	No effects were found
Yuan & Ellis (2003)	.5 min. no planning 10 min. pre-task plan. .5 on-line planning All: Look & Tell	Narrative (picture stories)	TOEFL 370-520	Analytic Measures	ANOVAs	Pre-task planning enhances complexity On-line planning influences accuracy & complexity

same from particular performance conditions, i.e. pre-task planning, or task characteristics, i.e. task structure.

The term 'language proficiency' has been long used in the context of language teaching and testing to refer to knowledge, competence and ability in the use of language. 'Language proficiency' has been used to represent various concepts such as communicative competence and communicative language proficiency. Commonly, however, 'proficiency' refers to a learner's general language ability in speaking, listening, reading and/or writing based on some kind of criteria or measure (Leeser, 2004). Likewise, in the present study 'language proficiency' is used to refer to the language ability of L2 learners or test-takers irrespective of how, where or under what conditions this ability has been acquired. Language proficiency as a construct has been a controversial issue over the past decades (See the detailed discussions of the construct of oral language ability in Chapter III). However, it is now generally agreed that language ability consists of several distinct but related constructs in addition to a general construct of language proficiency (Bachman, 1990).

Learners' language proficiency is usually categorized into different levels ranging from beginners to advanced or native-like proficiency levels. In practical situations, language proficiency tests are used to determine learners' proficiency levels. However, it has been constantly argued that language tests may fail to recognize the real distinction between language ability and the actual performance of ability (Bachman, 1990; Upshur, 1979).

A number of SLA studies have revealed that the nature of language proficiency might change at different stages of its development (Coady, 1979; Cohen, 1984; Farhady & Abbasian, 2001). It is also suggested that language learners at different levels of

language proficiency use various types of strategies that would affect their performance (Purpura, 1998). Young (1995) found a number of proficiency-related differences between the performances of advanced and intermediate learners. Clapham's study (1996) revealed that even the ability to use background knowledge requires a certain level of language proficiency.

In task-based research, it is suggested that one source of variability in the performance of L2 learners on tasks might be different levels of language proficiency (Robinson, 2001, Wigglesworth, 1997, 2001). Learners at different levels of language proficiency might act upon task characteristics and task conditions differently. Leeser (2004) has investigated whether the proficiency level of learners involved in a dictogloss task would influence their performance in terms of the number, type and outcome of their language-related episodes (when learners talk about or question their own language use). Results of his study indicate that higher-proficiency learners were more successful in performing the tasks and used more language-related episodes. Wigglesworth (1997) has investigated the effect of planning time on the performance of low-proficiency and high-proficiency test-takers. Results of her study reveal that test-takers of higher proficiency levels take advantage of planning time in a more efficient way. In fact, high-proficiency test-takers, compared to low-proficiency test-takers, benefited more from the presence of pre-task planning time on the more difficult tasks, with regard to the accuracy, complexity and fluency of their performance. Wigglesworth (2001) has also compared the effect of task structure, task familiarity and planning time on learners of high and low proficiency levels. Results of her study also suggest that planning time has been more helpful for the high proficiency test-takers performing the more cognitively demanding, i.e. unstructured, tasks.

The significance of language proficiency level in task-based research and the need for more research in this area is indicated in recent SLA literature (Robinson, 2001, Wigglesworth, 1997, 2001). However, not many studies have attempted to investigate the effect of different proficiency levels on language performance in task-based studies. Hence, the current study has attempted to explore whether different levels of language proficiency influence performance on tasks and whether they have any interactions with task characteristics and task conditions. The details of how language proficiency is incorporated into the current study will be presented in Chapter V.

4.5 Measuring Language Performance in the Present Study

Measuring second language performance has been the main purpose of a number of language teaching and most language testing studies. However, there are different ways in which oral performance on tasks can be measured. In general, there are two major methods of measuring oral language performance: rating procedures and analytic detailed measures. The detailed discussion and evaluation of each method of measuring oral performance is presented in the two sections that follow.

4.5.1 Rating Procedures

Evaluating language performance on the basis of rating procedures by some trained raters has been a tradition in language testing. By definition, rating procedures of this type of assessment refer to the " agreed procedures followed by raters in judging the quality of performances, particularly in the assessment of speaking and writing" (McNamara, 2000, p. 136). Extending back to the 1950s, a primary and commonly used type of rating scales of language performance was employed by the FSI (Foreign Service Institute) test which was initially adopted with the purpose of recruiting

personnel for official posts abroad. Since then, rating-mediated assessment has become more central to language testing. In such a framework, the criteria for recognizing performances of a given level are considered and then decisions about the number of the levels of performance are made (McNamara, 2000). In fact, language samples are assessed in terms of their quality and evaluated according to the relevant rating scales which describe different levels of language proficiency.

The rating scale systems normally differ in terms of the criteria they consider and the number of levels they recognize for each criterion. For instance, the FSI rating scale adopts a five-criterion scaling system of accent, grammar, vocabulary, fluency and comprehension (Manual of ETS, 1982). This system assumes six levels of proficiency on a continuum ranging from very low to highly advanced proficiency levels. The centre for Canadian Language Benchmarks (2000), in their recent rating scale manual, considers the following criteria for the assessment of L2 learners' speaking: accuracy of grammar, adequacy of vocabulary, intelligibility of speech, appropriateness and organization of discourse/coherence. These benchmarks are also defined at the three levels of elementary, intermediate and advanced proficiency, each with four sub-categories in terms of degree of proficiency. As a result, there are 12 proficiency-related speaking benchmarks ranging from "initial basic proficiency" at level 1 to "fluent advanced proficiency" at level 12 (Centre for Canadian Language Benchmark, 2000).

Some investigators, exploring the effect of task characteristics on language performance, have used this approach to performance ratings in the context of conventional methods of language testing (Wigglesworth, 2001; Elder et al, 2002). However, as discussed in Chapter III, the rating procedures of assessing language performance impose certain fundamental problems and disadvantages on the

assessment of tasks (See Chapter III for a detailed discussion). For this reason, the rating approach to assessing language performance is not frequently used in task-based studies.

4.5.2 Analytic Detailed Measures

A large number of researchers in the field of task-based studies have employed analytic detailed measures in assessing oral language performance. With regard to different aspects of performance, researchers have adopted various ways of operationalizing the measured performance. Skehan (2003) proposes that these different choices, to a large extent, are reflected in the theoretical positions researchers assume in carrying out their studies. He further explains that, for instance, researchers supporting a "Negotiation of Meaning" approach to task-based studies have used clarification requests, confirmation checks, comprehension checks and recast as the detailed measures they employ to assess performance on tasks (e.g. Long, 1989; Pica, 1994). Investigating task performance from a Sociocultural-theory perspective, some researchers have used measures of interactive involvement and measures of interactive symmetry (e.g. Duff, 1993; Van Lier & Matsu, 2000). And finally, researchers from a Cognitive approach have taken measures of fluency, accuracy, complexity and lexical density/variety to assess performance on tasks (Mehnert, 1998; Robinson, 2000).

Within the analytic detailed measures, there are two different methods of measuring performance: generalized and specific detailed measures. Some researchers have measured participant performance by employing a set of generalized measures, e.g. error-free utterances, whereas others have selected a number of specific measures, e.g. correct use of past tense or indefinite article. It is worth mentioning that a large body

of more recent task-based research has used generalized measures of fluency, accuracy and complexity to assess performance on tasks (Bygate, 1996, 2001; Ellis, 2002; Mehnert, 1998; Ortega, 1995, 1999; Robinson, 1995,2001; Wigglesworth, 1997, 2001). In contrast, some earlier researchers used different specific measures (Fotos and Ellis, 1991; Crookes, 1989). This group of researchers used specific measures either because they were using tasks with targeted specific structures or because they were testing hypotheses which were based on specific structures.

Foster and Skehan (1996) argue that generalized measures are more suitable for task performance studies because they are more sensitive indices of task performance. Drawing upon the contrast proposed by Widdowson (1989) between analyzability and accessibility, Foster and Skehan (1996) suggest the three measures of fluency, accuracy and complexity as the assessment criteria for task performance. Based on Widdowson's proposal, they conclude that fluency reflects the availability of learner accessible language. On the other hand, analyzability refers to the systematicity of interlanguage and the way in which it is organized so that rule-based performance may develop. In this context, Foster and Skehan (1996) argue that analyzability would include "attention to accuracy and a willingness to attempt ambitious forms" (p. 190). Skehan (2003) claims that "the complexity-accuracy-fluency dimensions of task performance have been justified both theoretically and empirically" (p. 22). He further argues that, theoretically, the sequence implies the three stages of: change in the underlying system, i.e. greater complexity; acquisition of greater control of the interlanguage system, i.e. greater accuracy; and development of performance control, i.e. fluency. In the following section I will describe each of the three measures of fluency, accuracy and complexity and will explain how different researchers have operationalized them in their studies.

4.5.2.1 Fluency. Fluency in SLA studies, in a very general sense of the term, refers to ease or automaticity in the learner speech and represents flow, continuity and smoothness of speech. Despite such a simplistic definition, fluency is a complex phenomenon that encompasses a multitude of linguistic, psycholinguistic and sociolinguistic features. Koponen and Riggenbach (2000) have discussed different aspects and representations of fluency in detail and mentioned that fluency includes a number of interconnected phenomena. They argue that fluency may refer to smoothness of speech in terms of temporal, phonetic, and acoustic features; it may represent proficiency at a macro or micro level; it may mean the automaticity of psychological processes; or it may be expressed as a notion contrasting the concept of accuracy. Fillmore (1979) also discussed the different definitions of fluency that were used by a number of researchers. He argued that:

> Fluency might simply be the ability to talk of length with few pauses; the ability to fill time with talk; the ability to talk in coherent and semantically dense sentences; the ability to have appropriate things to say in a wide range of contexts; and the ability to be creative and imaginative in the language use (Fillmore, 1979, p. 51).

Freed (2000) mentions that a survey of the construct of fluency reveals that explorations of the notions of fluency span a continuum that ranges from studies of its psychological manifestations and reflections of underlying speech-planning and thinking processes to studies of speech production, hesitation phenomena, and temporal dimensions of speech.

Based on the multifaceted nature of fluency, different researchers have adopted various measures to assess fluency. These measures, however, can be categorized into three sub-dimensions of fluency. The first sub-dimension of fluency is known as

silence, or as Skehan (2003) puts it, *breakdown* fluency. Length and amount of unfilled pauses, filled pauses and total amount of silence are some measures researchers use to assess *breakdown* fluency. Foster and Skehan (1996) and Skehan and Foster (1997) included in their analysis pauses of 1 second and over, and found that under planned conditions participants paused significantly less frequently. The amount of total pausing was also significantly smaller for planners. Mehnert (1998) also measured pauses of 1 second and longer as an indication of dysfluency.

Although a number of researchers have operationalized pausing through pauses of 1 second or more, recent SLA literature suggests that smaller amounts of pausing are better indicators of such fluency. Oppenheim (2000), following Stern (1992), proposes that pauses of less than half a second between short stretches of speech are one of the five characteristics of native-like delivery of American English. Freed (2000), in a study aimed at exploring the construct of fluency in the speech of L2 learners of French, investigated fluency in terms of 7 measures including unfilled pauses. Regarding the unfilled pauses, Freed measured the silences of longer than .4 a second that occurred at places other than predictable juncture boundaries. She argues that:

> Since silent pauses of shorter duration, frequently termed micropauses and measured in milliseconds, are characteristics of native speech and accurately measured by computerized acoustic analysis, we chose to identify and measure only those unfilled pauses [.4 a second or larger] that were heard as dysfluent and that usually did not occur at a clause boundary (Freed, 2000, p. 248).

A second sub-dimension is *speed fluency* and deals with the speed with which language is performed. Measures of speech rate, articulation rate, amount of speech, time ratio and mean length of run are usually used to show how fast language

performance is produced. Speech rate and length of run are the two commonly used measures of *speed fluency* in SLA studies. Mehnert (1998), Towell et al. (1996) and Freed (2000) have used mean length of run to measure fluency of speech production. Mean length of run in Towell et al (1996) is calculated as the mean number of syllables produced in utterances between pauses of .28 seconds and above. Mehnert (1998) found mean length of run by calculating the mean number of the syllables between pauses of 1 second. Freed (2000) defines length of run as continuous streams of running speech (measured in words) not interrupted by dysfluent pauses or hesitations. In effect, mean length of run is a manifestation of how lengthy the language produced between two pause boundaries is. Speech rate, i.e. number of syllables or words on average per minute, is another measure frequently used by researchers as an index of fluency (Ellis, 2002; Mehnert, 1998; Raupach, 1980; Robinson, 2001). Freed (2000) has measured speech rate on the basis of the number of "nonrepeated" words or semantic units per minute. Towell et al. (1996) have calculated speech rate by dividing the total number of syllables produced in a given speech sample by the amount of total time including the pauses. It can be concluded, thus, that speech rate refers to how fast and dense the produced language is in terms of the time units.

The third sub-dimension of fluency is what is known as *repair fluency* (Skehan, 2003). *Repair fluency* includes reformulation, replacement, false start and repetition of words or phrases. Wigglesworth (1997) measured the percentages of clauses containing self-repairs and reported that planned performance is significantly more fluent than unplanned performance. Skehan and Foster (1999) used repetitions, false starts, reformulations and replacements to measure fluency of language performance. Freed (2000) operationalized *repair fluency* in terms of repetition of exact words,

syllables or phrases, reformulations, false starts, corrections and partial repeat in the learner speech.

Unfortunately, these various conceptualizations of the nature of fluency have not been thoroughly investigated in task-based studies. In other words, fluency, which is now recognized as a multifaceted construct, has not been carefully investigated in task-based studies. Lack of enough research on the wide concept of fluency could present problems with comparing the findings of different studies. The reason for lack of comprehensive research studies on fluency might be the difficulty that is usually associated with measuring such a multifaceted construct. However, with the recent improvements of technology and a wider availability of software programs that enable researchers to use various measures of fluency, it is more convenient to measure different aspects of fluency. Hence, in order to have a more detailed and precise exploration of the nature of fluency and to know what effects different task characteristics would have on various aspects of fluency in task-based context, the present study has attempted to investigate a wide range of different aspects of fluency. The details of different fluency measures adopted in the current study and the relevant measuring procedures will be discussed in Chapter V.

4.5.2.2 Accuracy. With measures of accuracy, there is greater consensus among researchers in task-based studies. In some studies accuracy has been investigated through specific measures, such as past tense morphemes (Ellis, 1987) and plural *-s* (Crookes, 1989; Wigglesworth, 1997). Since many of these specific measures did not reveal any significant differences between different planning or task conditions, it was concluded that such measures are not sensitive to detecting differences between experimental conditions (Skehan and Foster, 1999). Hence,

many researchers have started using general measures of accuracy, such as percentage of error free clauses, or errors per 100 words. Foster and Skehan (1996) and Skehan and Foster (1999) have used the number of error free clauses divided by the total number of clauses to represent the percentage of accuracy. In both studies accuracy was enhanced as the planning time was provided to the learners and when task structure was present. Ortega (1999) has measured accuracy by means of targetlike use of analysis of two grammatical areas: morphology agreement of a noun and its modifiers (including possessives, adjectives and quantifiers), and use of the Spanish article system. She has further argued that the global measures have the disadvantage of being too broad to capture small changes in targetlike use since they combine multiple error types and obscure errors that might be important at a given level of development. In order to make up for such a disadvantage, Mehnert (1998) has used general measures of percentage of error-free clauses and the number of errors per 100 words as well as more specific measures of word order and lexical choice error. Interestingly, results of her study showed that accuracy improved with only 1 minute planning time but did not increase with a longer planning time. In Chapter V, I will discuss in detail how accuracy is measured in the current study.

4.5.2.3 Complexity. Complexity of performance, in task-based studies in general, refers to the organization of what is said with regard to the variety of syntactic patterning and subordination of the language output. Ortega (2003) defines syntactic complexity as the range of forms that surface in language production and the degree of sophistication of such forms. She argues that

> This construct is important in second language research because of the assumption that language development entails, among other processes, the

growth of an L2 learner's syntactic repertoire and her or his ability to use that repertoire appropriately in a variety of situations. (Ortega, 2003, p. 492)

Complexity is normally associated with the willingness to use a more elaborate language or to take risks in using more complex structures. Like accuracy, with complexity the focus of assessing performance is the attention that is given to form rather than to meaning. However, unlike accuracy, with complexity researchers have tended to employ a wider range of different units of analysis as the basis for measuring complexity.

Although use of subordination seems to be a recurrent measure for investigating syntactic complexity, some researchers have used other measures of complexity in task-based studies. Crookes (1989), for example, has included VP range as a measure of complexity. Foster and Skehan (1996) have employed non-simple present tense, use of modals and conditionals to measure syntactic complexity. Foster and Skehan (1999) report that measuring complexity through an index of subordination has gradually proved to be a reliable index that correlates with other measures of complexity. However, the unit of analyzing language in terms of this subordination index has been subject to change. Crookes (1989) has used s-nodes, either a simple independent clause or a dependent finite or non-finite clause, to analyze speech production and concluded that this provides a broader measure of complexity. Measures of T-units (Hunt, 1965) or C-units (Brock, 1986) have been later used by some researchers (Pica et al. 1989; Chaudron, 1988). T-units are reported to be the most popular unit for the analysis of both written and spoken data (Foster, Tonkyn and Wigglesworth, 2000). Hunt (1965) defined the T-unit as essentially a main clause plus any other clauses which are dependent upon it. It is also defined as the shortest possible unit into which a piece of discourse can be cut without leaving any sentence

fragments as residue. However, the use of T-units has been criticized since the "non-clausal structures" and "sentence fragments" may be included or excluded from the analysis. Due to the problems T-units impose on spoken data and following Brock (1986), Skehan and Foster (1997) and Mehnert (1998) used C-unit as the unit of analyzing subordination. The C-unit is defined as an utterance providing referential or pragmatic meaning, consisting of either a simple clause, or an independent sub-clause unit, together with subordinate clauses associated with either. As a result, a C-unit may be made up of one simple independent finite clause plus one or more dependent finite or non-finite clauses but the unit is mainly dependent on the semantic load of the utterance. However, it is not known how intonation patterns and syntactic structures might influence C-units. Hence, C-units are also criticized as they need other grammatical and intonational units to clearly determine the units and their boundaries.

Foster, Tonkyn and Wigglesworth (2000) have discussed the analysis of the spoken data in detail and emphasized that such analysis requires a principled way of dividing the transcribed data into units in order to assess features of accuracy and complexity. Identifying the shortcomings of measures like T-units and C-units, they have introduced the AS-unit (Analysis of Speech Unit) as a syntactic unit which is valid for spoken data. They provide a number of reasons to show that the AS-unit is more appropriate than other units used by researchers before. First, they argue that studies of pausing in native-speaker speech suggest that syntactic units are real units of planning with pauses happening at the syntactic unit boundaries, especially clause boundaries. They then discuss that the AS-unit allows analysis of speech units that are longer than a single clause. In this case, the intonation and pause features of speech are also taken into consideration and as a result multi-clause units are possible.

In fact they claim that Hunt's T-unit is included in the AS-unit for measuring the complexity of spoken language, but that it also allows for the inclusion of independent sub-clausal units, which are common in speech.

Foster et al. (2000) define the AS-unit as "a single speaker's utterance consisting of *an independent clause, or sub-clausal unit*, together with any *subordinate clause(s)* associated with either" (p. 365). In this definition, an independent clause will be minimally a clause including a finite verb. An independent sub-clausal unit will consist of: either one or more phrases which can be elaborated to a full clause by means of recovery of ellipted elements from the context of the discourse or situation. The definition of the AS-unit also includes minor utterances which are one class of "irregular sentences" or "non-sentences" identified by Quirk, Greenbaum, Leech, and Svartvik (1985). Furthermore, Foster et al. (2000) explain that "a subordinate clause will consist minimally of a finite or non-finite verb element plus at least one other clause element (Subject, Object, Complement or Adverbial)" (p. 366).

4.6 Perceptions of Task Difficulty

Despite a large number of studies that have focused on task difficulty in task-based language teaching and testing, little attention has been paid to how second language learners and test-takers perceive different aspects of task difficulty. Various researchers have discussed the effects of task difficulty on language performance. However, few have attempted to investigate how learners perceive, internalize and react to such difficulty. This lack of attention is justified in language testing research since test-taker reactions are not considered as central to the test validation process (Bachman, 1990; Elder et al., 2002). In effect, there is a general consensus among language testers that "test validation is more properly left to experts with relevant

training in test development and analysis" (Elder et al., p. 350). However, investigations of task difficulty and its effect on L2 learner performance cannot be considered as exhaustive if learner perceptions toward this difficulty are not taken into consideration. Their perceptions of task difficulty would enable researchers to expand their insights into the concept of task difficulty and establish a more reliable index of task difficulty to be used in the selection of tasks for different pedagogic or assessment purposes.

Learners' perceptions of teaching activities have frequently been investigated in language teaching contexts. Barkhuizen (1998) reports that learners' perceptions of language teaching and learning tasks have often surprised teachers. He recommends that teachers constantly monitor learners' perceptions and consider them in their planning and in their practising the target language. Graham (2004) has explored perceptions of English students towards learning French and how they view their level of achievement. While the learners of this study attribute their success to effort, high ability and effective learning strategies, they have cited low ability and task difficulty as the main reasons for lack of achievement in French.

Research in the language testing area has provided evidence that test takers have preferences for certain types of tests and that some tasks are perceived to be either easier or more interesting than others (Shohamy, 1982; Zeinder, 1990). Scott and Madsen (1983) showed that learners with low levels of proficiency rated oral interview tasks less favorably than did more proficient learners. Fulcher (1996) has employed questionnaires and interviews to investigate the reactions of test takers, including their perceptions of task difficulty, to three types of tasks: a picture description, an interview and a group discussion. Fulcher argues that test-takers can be sophisticated commentators on the test-taking experience and their perceptions of

task difficulty should be taken into consideration by test developers. Investigating learner perceptions of task-difficulty, Robinson (2001) reports that task difficulty significantly affects learner ratings of difficulty and stress in line with the increased amount of task difficulty. Participants of his study felt less confident but more interested in more cognitively demanding tasks. In contrast, Elder et al. (2002) in a study investigating the impact of performance conditions on perceptions of task difficulty in an assessment setting, report that test-taker perceptions of task difficulty did not generally correspond to the hypothesized difficulty of task conditions. They further argue that test-taker perceptions, therefore, can not be considered as systematic feedback, either as a test design or in organizing test validity arguments.

It is evident that, despite the significance of learner perceptions of task difficulty in task-based studies, this area has remained, to a great extent, unexplored. There is, in effect, a great amount of work required to explore language learners' perceptions of task difficulty. Therefore, one salient purpose of the current study would be to explore the retrospective perceptions of the participants of the study in terms of the difficulty of the task they are performing.

CHAPTER V

Research Design: Study One

5.1 Overview

In an attempt to uncover the effects of characteristics and conditions of oral narrative tasks on language performance, Study One is designed to investigate how degree of task structure, pre-task planning time and language proficiency level would influence language performance and test-takers perceptions of task difficulty in an assessment setting. As discussed in Chapter IV, task structure is defined in terms of problem-solution or schematic sequential structure. The performance conditions of a task are either planned or unplanned and the participants of the study belong to elementary or intermediate language proficiency levels. In this chapter, I will first present the research hypotheses formulated for this study. Then, I will explain the research design, the actual narrative tasks, planning conditions and the proficiency levels of the participants of the study. As test-takers perceptions of task difficulty have been investigated through retrospective questionnaires, the following section will explain the development and use of the questionnaires. A report of the pilot study, the participants of the main study and the setting in which the test has been carried out will then be presented. Finally, I will provide an account of the specific measures I have employed in coding the data and will discuss the data, the coding process, and other relevant issues.

5.2 Hypotheses: Study One

Hypothesis 1: This hypothesis deals with the effect of task structure on language performance and is presented in three sub-hypotheses:

Hypothesis 1a: Language performance in structured tasks would be more fluent than performance in unstructured tasks. This follows from Skehan and Foster (1999) and Wigglesworth (2001), who found that performance in structured tasks is more fluent than performance in unstructured tasks.

Hypothesis 1b: Language performance in structured tasks would be more accurate than the performance in unstructured tasks. This follows from Foster and Skehan (1999) and Wigglesworth (2001), who found that performance in structured tasks is more accurate than the performance in unstructured tasks.

Hypothesis 1c: Language performance in structured tasks would be less complex than performance in unstructured tasks (See Foster and Skehan, 1999).

Hypothesis 2: Language performance in structured tasks would be, as a function of degree of structure in the four tasks indicated in this study, progressively more fluent and accurate than the performance in unstructured tasks.

Hypothesis 3: This hypothesis deals with the effect of planning conditions on language performance and is presented in three sub-hypotheses:

Hypothesis 3a: Language performance under planned conditions would be more fluent than that produced under unplanned conditions. This follows from a series of research findings supporting the effect of pre-task planning on the fluency of language performance (Foster and Skehan, 1996; Mehnert, 1998; Ortega, 1999; Wigglesworth, 1997).

Hypothesis 3b: Language performance under planned conditions would be more accurate than that produced under unplanned conditions. This hypothesis is formed as a number of studies (Skehan and Foster, 1996; Mehnert, 1998; Robinson, 2001) found more accurate performances associated with planning.

Hypothesis 3c: Language performance under planned condition would be more complex than that produced under unplanned conditions. This follows from Mehnert (1998), Robinson (2001) and Wigglesworth (1997) that found planned performance was more complex than unplanned performance.

Hypothesis 4: The effect of planning would be, as a function of degree of structure mentioned in this study, progressively greater for the structured tasks with respect to fluency and accuracy but not greater for complexity.

Hypothesis 5: This hypothesis deals with the effect of the level of language proficiency on performance. It should be noted that level of language proficiency is included in this study as much to explore the interactive effects as to explore any main effect. This hypothesis is represented in three sub-hypotheses:

Hypothesis 5a: High proficiency test-takers would generally perform better, and particularly benefit more from pre-task planning, in terms of fluency of performance, than low-proficiency test-takers. This follows from Wigglesworth (1997) who found planning was an advantage for high-proficiency test-takers.

Hypothesis 5b: High proficiency test-takers would generally perform better, and particularly benefit more from the pre-task planning, than low-proficiency test-takers in terms of the accuracy of their performance. This follows from

Wigglesworth (1997) who found that planning would help high-proficiency test-takers produce more accurate language.

Hypothesis 5c: High proficiency test-takers would generally perform better, and particularly benefit more from the pre-task planning, than low-proficiency test-takers in terms of the complexity of their performance. This follows from Wigglesworth (1997) who found that planning would allow high-proficiency test-takers produce more complex language.

Hypothesis 6: The performance of high-proficiency test-takers would benefit from planning more on the unstructured tasks than on the structured tasks with respect to fluency, accuracy and complexity. This follows from Wigglesworth (1997) who reported that high proficiency test-takers produce more fluent, accurate and complex language on difficult tasks rather than on easier tasks when they are provided with the pre-task planning.

Hypothesis 7: Test-taker perceptions of task difficulty are in line with the predicted difficulty of the tasks, in terms of task structure, in this study .This follows from Robinson (2001) who reported that learners rated cognitively complex tasks as significantly more stressful than simple tasks.

5.3 Methodology

5.3.1 Design

As there are three independent variables – task structure, pre-task planning and language proficiency- to be investigated in the present study, a factorial design is required in which "the effects of several independent variables may be tested at the same time" (Seliger &

Shohamy, 1989). Therefore, A 2 x 2 x 4 factorial design was used in the current study with planning condition, proficiency level, and task structure as the independent variables. Planning condition and language proficiency were between-participants variables and each had two levels with the participants belonging to either of the two conditions and levels. Task structure, which was operationalized through 4 different picture stories, had four levels representing a scale in the degree of structure of the tasks. Task structure was a within-participant variable and therefore, all participants performed on all the four levels, i.e. tasks. The dependent variables of the study were language performance represented through fluency, accuracy and complexity and test-takers perceptions of task difficulty measured by retrospective questionnaires.

5.3.2 Tasks

As discussed earlier, oral narrative tasks are frequently employed in the context of assessing second language performance (Elder, et al., 2002; Iwashita et al, 2001; Robinson, 2001). Oral narrative tasks are also routinely used as a single type of stimulus in eliciting language samples by international testing organisations (e.g. Test of Spoken English). Oral narrative tasks in this sense refer to stories based on a sequenced set of picture prompts which are given to participants/test-takers to elicit oral language performance. 'Task' in LT, as discussed in Chapter III, is usually used to denote a broader concept. However, in the context of task-based research and also in the study reported here, 'task' is sometimes used to refer to the actual picture stories that are employed to elicit oral language performance. As discussed in Chapter III, the rationale for using narratives is, to a great extent, justified in terms of construct validity, reliability

and authenticity of the test. However, the prime reason for selecting oral narrative tasks in the present study is to have conformity with the literature from which the theoretical assumptions of the study are drawn.

In order to find proper picture stories that suit the purpose of the study, two main sources were searched in detail: 1) EFL sources including course books and supplementary materials for teaching English and other modern languages; and 2) non-EFL sources including a wide range of different materials such as cartoon books, children's story books, newspapers and pictorial stories. The selection criteria were set to find picture series which were clear, had worthwhile stories to be told, were of a reasonable length suitable for the study, were culturally familiar to the participants, were neither linguistically cued nor linguistically demanding, and seemed interesting to the participants. A total of 25 picture stories initially seemed to meet all the criteria. However, because of cultural issues and practicality restrictions of the study, 7 picture stories were later excluded from the collection.

The remaining 18 picture series were carefully analyzed by two experienced researchers[1] and myself. Based on the given definitions of structure discussed in Chapter IV, the 18 picture stories were categorized into two groups of structured and unstructured tasks. In further discussions and analyses of the degree of structure in the picture stories and based on two notions of structure, the structured tasks were then categorized into a problem-solution and schematic sequential structure.

The unstructured picture stories were also categorized into two levels of less structured

---

[1]I am grateful to Peter Skehan and Constant Leung for spending many hours studying each of the picture stories very carefully to help me choose the ones that were the most suitable for the purpose of the study.

and least structured. As discussed in Chapter IV, the unstructured tasks were not based on either a problem-solution or a schematic sequential structure. They did not have a clear time line and the sequence of the events was arbitrary. However, with the unstructured picture stories the criterion for placing them in a less or more unstructured category was the number of pictures in each story that could be rearranged without the main theme of the story being compromised. In other words, a picture series with more pictures easily moveable and interchangeable with others is considered less structured than a picture series with fewer moveable pictures.

To achieve the purpose of the study regarding the degree of structure, two structured and two unstructured tasks were selected. From among the structured tasks, one picture series was selected to represent the problem-solution structure and one picture series was selected to represent the schematic sequential structure. As discussed in Chapter 4, a problem-solution structure is hypothesized to represent a higher degree of structure, than the schematic sequential one. Therefore, two structured tasks, a problem-solution and a schematic sequential, were required to indicate the two degrees of structure. The task selected from the problem-solution category, i.e. the Football picture story (Heaton, 1996), was a picture story with a transparent problem-solution structure and a well-presented sequential organization. The second structured task, Picnic (Heaton, 1966), on the other hand, was based on a clear schematic sequential organization and contained an implicitly stated problem which was only revealed in its last frame. However, this task did not propose a transparent problem-solution structure, which made it less structured than the Football task.

Similarly, two tasks with varying degrees of structure were selected from the unstructured category. A lack of a problem-solution relationship and an inadequately clear sequential organization suggested that both tasks were unstructured. However, they differed from one another for the amount of sequential organization they contained. The least structured task, i.e. Walkman (Swan & Walter, 1990), did not propose a clear time line or any sequential organization and, therefore, was less structured than the Unlucky Man (Ur, 1984), which had a loosely presented sequential organization. In other words, events in the Walkman task were arbitrarily related to one another and the sequence of organization of events hardly followed a timeline.

As mentioned before, task structure was operationalized in terms of the number of pictures that could be rearranged in each picture story without the main theme of the story being compromised. That is, the two unstructured tasks had pictures that could be simply rearranged with no real change in the story. However, the number of such movable pictures determined which task was the least structured. As in the Walkman task there were more pictures that could be rearranged, it was assumed to be the least structured task. With Unlucky Man, there was one group of related pictures that could be rearranged, whereas in Walkman, all the pictures, except for the first and the last pictures could be rearranged without any compromise in the theme of the story.

All the picture stories consisted of six pictures, except for the Unlucky Man task which had a set of moveable pictures in the middle and, therefore, had ten pictures. The four picture stories can be seen in Appendix 1. Figure 5.1 demonstrates how the above-mentioned four tasks can be located on a continuum representing a scale of the degree of structure hypothesized in the present study.

Figure 5.1

Degree of Structure in the Four Tasks: Study One

+ Structured			- Structured
1. Football	2. Picnic	3. Unlucky Man	4. Walkman

In order to avoid any results emerging from a practice effect, a counterbalanced design[2] was considered for the participants' performing the tasks. In effect, all the participants performed the four tasks but 25% of them started with the Football task, 25% with the Picnic task, 25% with the Unlucky Man and 25% with the Walkman task. Table 5.1 shows the four sequences of tasks in which the participants performed the tasks in this study.

Table 5.1

Counterbalanced Sequence of the Tasks: Study One

Sequence 1	Football	Picnic	Unlucky Man	Walkman
Sequence 2	Picnic	Unlucky Man	Walkman	Football
Sequence 3	Unlucky Man	Walkman	Football	Picnic
Sequence 4	Walkman	Football	Picnic	Unlucky Man

5.3.3 Planning Conditions

Pre-task planning time, in the present study, was operationalized in terms of the amount of planning time provided to the participants. In effect, there were two groups of

---

[2]It is a design in which the participants take different parts of a test in different orders. This design is used to minimize the possibility of a practice effect.

participants in terms of the planning time, i.e. test-takers who had time 5 minutes to plan before performing the tasks and those who had 30 seconds. As discussed earlier in Chapter IV, the amount of time given to the test-takers was influenced by the findings of previous research studies (Elder et al. 2002; Mehnert, 1998; Wigglesworth, 2001). The unplanned group, thus, was given 30 seconds to look at each of the picture stories and prepare themselves before they started telling the stories. It was considered that 30 seconds was too short for any planning to take place. Participants in the planned groups, on the other hand, were given 5 minutes to look at each of the picture stories and were advised to plan for telling the story. Moreover, under the planned conditions each participant was given a sheet of paper to take notes or to write what they wanted. However, they were informed that they would not be allowed to use their notes while they were telling the story. The instructions given to both groups were identical in all other regards. Planning was not operationalized in terms of the type or source of planning as the study was carried out in an assessment setting.

5.3.4 Language Proficiency Levels

The participants were drawn from two levels of language proficiency, i.e. elementary and intermediate. Prior to the study the participants were placed in their levels on the basis of an institutional placement test for language proficiency. This institutional test is locally developed and includes a number of different parts assessing students' ability in reading, writing, listening, speaking and language use. The reliability and validity of the tests are regularly checked. All participants of the study had taken and passed such a test two to three weeks before the study was conducted and the results of the institutional test were

taken as the main criteria in recruiting the participants at the two levels of proficiency. Nevertheless, to confirm the homogeneity of the groups and also to distinguish between the two proficiency levels, their language proficiency was tested by the "Oxford Placement Test 2" (Allan, 1992) (See Appendix 2 for a sample of the test). It should be noted that due to practical reasons, only the grammar part of the test was run. However, the results of their institutional test contained a speaking section which demonstrated that the participants had achieved a certain proficiency level in their oral language abilities. The participants' responses to Oxford Placement Test were checked and scored on a scale of 100 points. The elementary group had a range of scores between 17 to 44 and the intermediate group scored from 45 to 75. This ranking was mapped to band 3, relating to elementary and post elementary levels, and band 4, relating to pre and lower intermediate levels, defined in the Oxford Placement Test. For the fluency of discussions in the current and following chapters, however, bands 3 and 4 will be called elementary and intermediate levels respectively. These results were used as a second criterion for classifying the participants into elementary or intermediate levels. It is worth mentioning that a relatively large correlation (r = .56) was observed between the participants' institutional test and the Oxford Placement Test.

5.3.5 Perceptions of Task Difficulty

In order to explore participant perceptions of task difficulty, a questionnaire was designed and developed for each the planning condition. All participants were asked to complete a relevant questionnaire as soon as they had all four tasks. Both questionnaires contained questions about the participant's perceptions of task difficulty of the four tasks, and an

open-ended question for the participants to mention their suggestions and comments about the tasks. The planned-group questionnaire differed from the unplanned-group since it included a third question about the usefulness of the planning time for each of the four tasks. In effect, the participants in the planned groups were asked to comment on how useful they found the planning time in performing each of the tasks. Regarding task difficulty, answers were given on a four-point scale with 1 representing "very easy" tasks and 4 "very difficult" tasks. The answers to the extra question for the planned group were also given on a four-point scale with 1 indicating that the planning time "helped very much" and 4 showing that the planning time "did not help at all". To avoid any potential confusion or misinterpretation resulting from the participants' reading ability, the questionnaires were translated into the participants' first language. English versions of both the planned and unplanned questionnaires are provided in Appendix 3.

5.3.6 Pilot Study

In order to find out if the selected tasks were functioning in line with the theoretical assumptions of the study and to investigate whether there are features in the picture stories which might intrude on performance or cause differences, which in terms of the underlying principles are spurious, the four tasks were first pre-piloted and then piloted with 17 participants. In the pre-pilot stage, three Farsi-speakers, one at elementary level and two at intermediate level performed four tasks and completed the questionnaires. The results of the pre-pilot study revealed that one of the tasks seemed to be confusing to the participants. Therefore, after careful consideration of the other options, this task was replaced by another task with similar characteristics and the same type of structure.

The new set of four tasks and the two planning conditions were then piloted on 14 language learners of elementary and intermediate proficiency level in a college in London. Participants of the pilot study were aged between 18 and 24 and were from 3 different language backgrounds including 6 Farsi speakers, 7 Chinese speakers and 1 Arabic speaker. Although the main study was designed to investigate the effects of task characteristics on language performance of Iranian test-takers of English, the participants of the pilot study were drawn from different language backgrounds due to access restrictions of carrying the pilot study in London. Every effort was made to include as many Farsi speakers as possible in both the pre-pilot and the pilot studies. It is worth noting that the results of the pilot study did not reveal any first language-related differences among the performances of the different participants.

In the pilot study, the participants were assigned to either a planned or an unplanned condition and performed the tasks in a one-to-one setting with the researcher. They all performed the four tasks in a counterbalanced design and completed the questionnaires afterwards. The results of the pilot study suggested that the selection of tasks and the amount of planning time were appropriate, particularly in terms of the practical aspects of a testing setting. However, regarding the questionnaires, the results showed that the extra question for the planned group was problematic since it did not elicit differentiated answers. Therefore, this question was reconsidered and reworded so that it could be used in the main study.

.5.3.7 Participants in the Main Study

The participants in the main study were 80 language learners studying English at Simin Educational Association[3] in Tehran, Iran.. They were all adult females aged between 18 and 45. They were studying English as a foreign language at an elementary or intermediate level and had been studying English at the same language school for at least 18 months. The participants were Farsi speakers and had a similar language learning history both within the public schooling system and at the above-mentioned language school. But they differed regarding the period of time they had been studying English in the past, the contact they had with English outside classroom and the purposes for which they were studying English.

As they had already taken part in similar testing situations in their language school and had performed similar tasks, they were all familiar with both the testing conditions and the test format, i.e. oral narratives. Before performing the oral narrative tasks, they were briefed on the purpose of the study and were asked to take the Oxford Placement Test.

One participant was withdrawn from the study and replaced with another one as she expressed her unwillingness in the middle of the test. The participants at each language proficiency level were randomly assigned to a planned or unplanned condition and one of the four sequences of the counterbalanced design demonstrated in Table 5.2.

---

[3]The above-mentioned institution has given consent to their name being reported in this study and any relevant presentations and publications.

Table 5.2

Design of Study One

Planning condition	*Proficiency Level*	*No. of Participants In Sequence1*	*No. of Participants In Sequence2*	*No. of Participants in Sequence 3*	*No. of Participants In Sequence 4*
Planned	Low-proficiency	5	5	5	5
	High-proficiency	5	5	5	5
Unplanned	Low-proficiency	5	5	5	5
	High-proficiency	5	5	5	5

5.3.8 Setting of Administration

As discussed earlier, all the participants were tested by an" Oxford Placement Test" before they performed the four oral narrative tasks. Some participants were tested two or three days before they performed the oral narrative tasks but others had the Oxford test on the same day they had the oral narrative tasks. In order to collect data in an assessment setting, every effort was made to make sure the test is administered in a very similar way to the assessment settings usually created both in TBA studies (e.g. Elder et al., 2002) and in their institutional testing context. To elicit the participants' performance on the oral narrative tasks, all participants were tested in a one-to-one setting with the researcher. The researcher met the participants individually and explained the purpose of the test to them. Each participant was randomly assigned to either the unplanned or planned conditions and to one of the four counterbalanced sequences of the four tasks.

Then, the instructions of the task were given to the participants. The participants were given each of the picture stories one at a time. Under the unplanned condition, they were told that they had just 30 seconds to look at each picture story before they should start

telling the story to the researcher in such a way that the researcher who did not have access to the stories could understand what was happening in each story. After the initial 30 seconds, the participants had the picture stories in front of them and were able to look at them while they were telling the story to the researcher who tape-recorded the participant's performance on the first task. For each of the tasks the participants were given 3 to 4 minutes time to tell the story. Then, the same procedures were taken for the second, third and fourth task, one after the other.

Under the planned conditions, the participants were told that they had 5 minutes to look at each picture story and plan for what to say and how to tell the story. They were also given some paper to take notes if they wished. But they were informed that they would not be allowed to use their notes while they were telling the stories. They were reminded that they should tell the story in a way that the researcher who did not have access to the stories could understand what was happening in each story. Then they were given 3 to 4 minutes to tell the story. After 5 minutes, the participants had the picture stories in front of them and were able to look at them while they were telling the story to the researcher who tape-recorded the participant's performance on the first task. Then, the same procedures were taken for the second, third and fourth task, one after the other. The notes participants made under the planned conditions were all collected and kept for a further analysis. Under both conditions, the main criterion considered for the participant performance was completion of each of the tasks. In other words, if they were able to complete the task they were counted as a participant of the study. Otherwise, they were substituted with another participant.

Based on the type of planning conditions, the participants were asked to complete a relevant questionnaire. They were also encouraged to comment about the test and the tasks in general in the last question of the questionnaire. To avoid any misunderstanding on the side of the participants, all the introductory talk and the instructions to the participants were given in Farsi. Table 5.2 demonstrates the design of the study in terms of the planning conditions, proficiency levels and the task sequences.

5.4 The Analytic Detailed Measures Adopted in This Study

As discussed earlier in Chapters III and IV, I decided to employ analytic detailed measures to investigate the three aspects of language performance, i.e. fluency, accuracy and complexity, in the present study. A number of different measures are adopted to investigate fluency, one measure for accuracy and one measure for complexity. These measures are used to analyze performance and to find out whether performance would vary as a function of task structure, planning conditions and level of language proficiency. The section that follows will provide a detailed description of how fluency, accuracy and complexity are measured.

5.4.1 Fluency Measures

There are a number of different measures of fluency which are generally assumed as significant indicators of fluency. Drawing on SLA and LT literature, in the present study fluency is measured through the number of false starts, reformulations, replacements, repetitions, length of run, speech rate, number of pauses, mean length of pauses, proportion of time spoken and total amount of silence.

Temporal aspects of fluency including pauses and total amount of speaking time were measured digitally. In fact, as the manual methods of measuring pauses are not very accurate and are generally subject to some degree of measurement error, it was decided that the pauses and total amount of speaking time be measured digitally. To achieve such a purpose and to avoid measurement errors the Goldwave software[4] (Digital Audio Editor, 2001) was used. Pauses of longer than .4 of a second were measured and the relevant codings were inserted in the transcribed data. However, filled pauses such as *em, er, uh,* and *eh* were not measured in this case but were simply indicated in the transcribed data. The speech rate was calculated by dividing the total number of syllables produced in each performance by the total amount of time expressed in seconds. The mean length of run was calculated by finding the mean number of syllables produced in utterances between breaks of .4 of a second and above. The proportion of time spoken was calculated by finding the mean length of time a participant actually speaks, excluding the pauses and silences, in each performance. As pauses of longer than .4 of a second were measured, mean length of such pauses and total length of silence for each task was then calculated.

Repair fluency included a number of different measures. False starts referred to all utterances that were abandoned before completion (example: and some time # after a few minutes). Reformulations were identified as phrases or clauses that are repeated with some modifications in their syntax, morphology or word order (example: they understood # they understand ~). Replacements were identified as lexical items that were

---

[4]Goldwave is a digital sound editor, player recorder and converter. Using Goldwave software, one can play and select any part of a sound. It is also possible to measure any duration of a sound or silence.

substituted for another (example: they started that # they noted that rpl). It should be noted that reformulations and replacements are always preceded by a false start. Repetitions were identified as the immediate and verbatim repetition of words or phrases (example: the mother is making making * tea). All these measures were coded and represented in the transcribed data using specific symbols. For a list of the coding symbols see Appendix 4 and to see samples of the transcribed and coded data refer to Appendix 5.

5.4.2 Complexity Measure

Following Foster et al. (2000), the transcribed data was divided into AS units and dependent clauses. As discussed earlier in Chapter IV, AS units are more valid for measuring spoken data since they are syntactic units which allow the intonation and pause features of speech to be taken into consideration (Foster et al., 2000). Hence, the data in the present study were divided into AS units that contained independent clauses, subordinate clauses and sub-clausal units. The intonation and pausing patterns of speech had a direct influence on determining whether a clause was an independent clause or a dependent one. As a result, the complexity of the performance was measured through an index of subordination by dividing the number of clauses by the number of AS units.

5.4.3 Accuracy Measure

Accuracy, in the current study, was measured in terms of an index of error-free clauses. Error-free clauses were defined as clauses in which no error was seen with regard to syntax, morphology, native-like lexical choice or word order. Errors in stress, intonation patterns or pronunciation of the words and utterances were not included. The native-like

use of the language, in terms of grammar and lexis, was generally considered as a criterion in determining whether the clauses were error-free. All error-free clauses were then identified and coded in the transcribed data. The ratio of error-free clauses to the total number of clauses was the general measure of accuracy employed in this study. A computer program that is specifically designed to analyze language performance data was used to calculate the ratio of error-free clauses and the index of subordination.

5.5 Data

The recorded performances of all 80 participants of the study were transcribed, using a technical transcribing machine, and were then word-processed. Using the Goldwave software (2001), all the tape-recorded data were digitized and transferred to CDs so that they were compatible with the use of computer software in measuring different aspects of language performance. The transcription of all performances were coded in a unified systematic way by inserting the specific coding symbols that represented different measures of fluency, accuracy and complexity (as explained in the previous section). The transcribed and coded data were continuously checked with the recorded audio performances throughout the coding and analyzing phases of the study. Since the intonation pattern played an important role in identifying the AS unit, it was necessary to compare the recorded data with the coded transcripts of the data. Various analyses were then carried out on the performance of each participant and each task to determine the values of the dependent variables, i.e. different measures of fluency, accuracy and complexity.

5.5.1 Coding the Data and Inter-Rater Reliability

Before coding the data, I received training in coding language performance data from an experienced researcher[5] specialized in this area. After the initial training, I received further training in coding measures of accuracy and complexity and where problems arose. The experienced researcher then coded 10% of the data against which the data coded by me were tested. Table 5.3 shows the results of the final inter-rater reliability. A reliability coefficient of higher than 90% was achieved in coding the complexity measures, i.e. the AS units and the dependent clauses, repetitions and replacements. However, the reliability coefficient for measures of accuracy, false starts and reformulations were initially lower. With further training and through careful considerations of the measures of accuracy, false starts and reformulations all the data were reviewed and re-coded until a higher correlation was achieved for all measures.

Table 5.3

Inter-Rater Reliability Coefficient for the Coded Data

Measures	*AS Unit*	*Dependent Clause*	*Error-free Clause*	*Reformulation*	*Replacement*	*Repetition*	*False Start*
Pearson-pro. Coefficient	.99	.98	.94	.97	1.00	1.00	.96

5.5.2 Computer Program Used to Analyze the Data

A computer program (work-in-progress) was especially adapted for this study. This program was used to read the coded data and to compute different measures of fluency,

---

[5]Many thanks to Pauline Foster for training me how to code the transcribed language performance data, for patiently helping me with my endless list of questions and above all for accepting to code 10% of the data as a second rater.

complexity and accuracy. After the data were coded and values of the 12 measures for each of the tasks were established, the resulting complex data set was subjected to a series of statistical tests to determine whether the independent variables of task structure, planning time and proficiency level had any statistically significant effect on different aspects of the participant language performance. The detailed statistical analyses and the results obtained from the various analyses are presented in the Chapter VI.

CHAPTER VI

Analyses and Results: Study One

6.1 Introduction

As discussed earlier, a 2 x 2 x 4 factorial design is considered as the research design of the study reported here. There are three independent variables in the present study: task structure, pre-task planning and proficiency level. Task structure is a within-participant variable being represented at four levels to indicate varying degree of structure. Pre-task planning and proficiency level are both between-participant variables, each with two levels. In total, 12 dependent variables are used to measure different aspects of fluency, accuracy and complexity of the oral performance. Retrospective questionnaires are also employed to explore how the test-takers perceive task difficulty as a function of task structure, planning time and language proficiency. As mentioned in Chapter V, with the obtained figures for all 12 measures of fluency, accuracy and complexity, a complex database file was made, which included the results of the detailed performance of all 80 participants on different tasks, planning conditions and proficiency levels. Furthermore, a separate database file was made for the responses the test-takers provided to the questionnaires on perceptions of task difficulty.

A wide range of statistical analyses was then employed, using SPSS 9.0 for Windows, to test different hypotheses of the study. In order to uncover whether the measures in these three sets of variables, i.e. fluency, accuracy and complexity, were related to one another or whether they truly represented three distinct factors, the data were initially

subjected to a factor analysis for each individual task. Based on the results of the factor analyses, a repeated measures MANOVA was then performed to test the overall effect of task structure, planning time and proficiency level on language performance. Finally a series of ANOVAs and t-tests were run to explore the differences among the tasks, between the planning conditions and language proficiency levels as well as the interactions among task structure, planning time, and proficiency level. Finally, to investigate whether task structure, planning time and proficiency level had any effect on test-takers' perceptions of task difficulty, a three-way ANOVA was conducted.

6.2 Statistical Analyses

6.2.1 Underlying Factors in Language Performance

A separate factor analysis was run for each of the four tasks with the 12 measures of reformulation, false start, replacement, repetition, accuracy, complexity, length of run, speech rate, total amount of silence, proportion of time spoken, number of pauses and mean length of pauses (see pages 159 and 160 for the relevant data). Prior to performing the analysis, the suitability of the data for factor analysis was investigated. Inspection of the correlation matrices revealed that there were many coefficients of .4 and above in each matrix. The Kaiser-Meyer-Oklin values were above .68 for all the tasks, exceeding the recommended value of .60 (Kaiser, 1974). Barlett's Test of Sphericity reached statistical significance, supporting the factorability of each correlation matrix. All the evidence confirmed the suitability of the data for factor analysis. The results shown in Tables 6.1 to 6.4 demonstrate all factor loadings of .40 and above, and the communality scores of the measures which indicate the amount of variance that is explained by the factor. Interestingly, three factors with a fixed

number of measures were extracted from each of the analyses. Discussions of the factors and the measures will follow.

Table 6.1
Factor Analysis for the Walkman Task

Measures	Factor 1	Factor 2	Factor 3	Communality
Reformulations		.83		.732
False starts		.93		.865
Replacements		.60		.529
Repetitions		.66		.462
Accuracy			.65	.511
Complexity			.82	.701
Length of run	.65			.709
Speech rate	.85			.741
Total silence	-.89			.857
Prop. time spoken	.88			.830
No. of pauses	-.46	-.48		.500
Mean length of pause	-.89			.812

Table 6.2
Factor Analysis for the Unlucky Man Task

Measures	Factor 1	Factor 2	Factor 3	Communality
Reformulations		.81		.739
False starts		.90		.865
Replacements		.44	.60	.604
Repetitions		.60		.396
Accuracy			.54	.493
Complexity			.68	.505
Length of run	.63			.681
Speech rate	.84			.771
Total silence	-.87			.805
Prop. time spoken	.92			.852
No. of pauses	-.66			.675
Mean length of pause	-.92			.857

Table 6.3
Factor Analysis for the Picnic Task

Measures	Factor 1	Factor 2	Factor 3	Communality
Reformulations		.86		.791
False starts		.92		.867
Replacements		.59		.353
Repetitions		.68		.579
Accuracy			.72	.530
Complexity			.79	.648
Length of run	-.43		.72	.732
Speech rate	-.44		.60	.595
Total silence	.92			.809
Prop. time spoken	-.90			.903
No. of pauses	.88			.808
Mean length of pause	.82			.848

Table 6.4

Factor Analysis for the Football Task

Measures	Factor 1	Factor 2	Factor 3	Communality
Reformulations		.88		.880
False starts		.94		.892
Replacements		.41		.276
Repetitions		.62		.490
Accuracy			.65	.662
Complexity			.87	.716
Length of run	-.66	-.44	.43	.767
Speech rate	-.84			.793
Total silence	.95			.912
Prop. time spoken	-.94			.902
No. of pauses	.80			.736
Mean length of pause	.87			.844

As the results of the factor analyses reveal, for all the four tasks, Factor 1 is made up of six high loadings on the same measures of fluency. These measures are length of run, speech rate, total amount of silence, proportion of time spoken, number of pauses and length of pauses. Comprising a significant loading factor, these measures refer to different temporal aspects of fluency and suggest that fluency affect all these different measures to the same extent. Length of run and speech rate for the Picnic task load on Factor 1 but they have a higher loading on Factor 3 suggesting that participants with a higher length of run and speech rate have achieved more accuracy and complexity in their performance. Although the main loading on length of run is on Factor 1, for Football there is also a slight loading on Factor 3 indicating an association between length of run and measures of fluency and accuracy. For the Walkman task, number of pauses loads on Factor 2 as well as Factor 1. Nonetheless, all these loadings confirm the existence of one general temporal fluency factor in the data. This hypothetically means that the more fluent participants would be expected to use a significantly higher length of run, a faster speech rate, less amount of silence, a fewer number of pauses, shorter pauses as well as more time spent speaking during their

performance. The results of the factor analysis in Mehnert's study (2001) support the same loadings for speech rate, length of run and total amount of silence.

The second factor loading is based on reformulations, false starts, replacements and repetitions. Principally considered as indicators of fluency, these measures tend to load highly on a second factor representing another aspect of fluency. Since these processes are involved when candidates attempt to repair their performances, this aspect of fluency is frequently called repair fluency in SLA literature (Freed, 2000; Skehan, 2001, 2003). The loadings for reformulation and false starts define the factor with higher loadings, while replacement and repetition follow them with lower, yet significant loadings. Replacement loads more on Factor 3, i.e. along accuracy and complexity, for the Unlucky Man task, suggesting that more replacements are associated with more accurate and more complex performances. However, in all other cases the high loadings of the four measures are located on Factor 2.

As discussed in Chapter V, the ratio of error-free clauses in each performance represents the accuracy measure and the ratio of subordination of each performance accounts for the complexity measure. The results of the factor analyses reveal that accuracy and complexity load highly on a third factor suggesting that more accurate language was also more complex. Furthermore, the fact that these measures are associated with each other indicates that they are likely to reflect the same underlying constructs. This confirms the idea that accuracy and complexity are both aspects of form and are in contrast with fluency which is assumed to be an aspect of meaning.

The loading of accuracy and complexity on the same factor also suggests that the two measures have a certain degree of association or go-togetherness.

Interestingly, the loadings indicate that the four factor analyses were very stable across all the tasks. In fact, with all the four factor analyses, the same number of

factors and the same order of factors and loadings are given for all the tasks. Factor one consistently loads on temporal fluency measures; Factor 2 loads on repair fluency measures; and Factor 3 loads on measures of form, i.e. accuracy and complexity. As part of factor analyses, the correlation matrices for all four tasks are presented in Tables 6.5 to 6.8.

Table 6.5

Correlation Matrix for the Walkman Task

Correl	reform	falstar	replac	repetit	accura	Comp x	lofrun	spchrt	nofpas	totsiln	timspk	menps
reform	1.00	.865	.274	.334	-.192	.012	-.242	.026	.280	.169	.023	-.057
falstar	.865	1.00	.516	.453	-.144	.077	-.257	-.053	.351	.187	.065	-.057
replac	.274	.516	1.00	.304	-.030	.186	-.150	-.079	.142	.167	.021	.017
repetit	.334	.453	.304	1.00	-.055	.068	-.371	-.216	.384	.173	.036	.004
accura	-.192	-.144	-.030	-.055	1.00	.318	.467	.435	-.175	-.312	.377	-.301
compx	.012	.077	.186	.068	.318	1.00	.302	.170	-.191	-.246	.290	-.147
lofrun	-.242	-.257	-.150	-.371	.467	.302	1.00	.712	-.538	-.593	.585	-.420
spchrt	.026	-.053	-.179	-.216	.435	.170	.712	1.00	-.418	-.696	.608	-.698
nofpas	.280	.351	.142	.384	-.175	-.191	-.538	-.418	1.00	.609	-.409	.160
totsiln	.169	.187	.167	.173	-.312	-.246	-.593	-.696	.609	1.00	-.823	.784
timspk	.023	.065	.021	.036	.377	.290	.585	.308	-.409	-.823	1.00	-.820
menps	-.057	-.057	.017	.004	-.301	-.147	-.420	-.698	.160	.784	-.820	1.00

Table 6.6

Correlation Matrix for the Unlucky Man Task

Correl	reform	falstar	replac	repetit	accura	Comp x	lofrun	spchrt	nofpas	totsiln	timspk	menps
reform	1.00	.740	.012	.313	-.293	-.060	-.245	-.024	.343	.159	-.009	-.025
falstar	.740	1.00	.516	.389	-.168	.036	-.274	-.039	.356	.215	-.069	.080
replac	.012	.516	1.00	.180	-.018	.143	-.087	.039	.052	.009	.092	.001
repetit	.313	.389	.180	1.00	-.176	-.028	-.348	-.124	.435	.312	-.177	.093
accura	-.293	-.168	-.018	-.176	1.00	.214	.463	.322	-.312	-.293	.275	-.191
compx	-.060	.036	.143	-.028	.214	1.00	.339	.310	-.161	-.203	.190	-.199
lofrun	-.245	-.274	-.087	-.348	.463	.339	1.00	.655	-.568	-.542	.598	-.477
spchrt	-.024	-.039	.039	-.124	.322	.310	.655	1.00	-.580	-.700	.697	-.731
nofpas	.343	.356	.052	.435	-.312	-.161	-.568	-.580	1.00	.782	-.601	.373
totsiln	.159	.215	.009	.312	-.293	-.203	-.542	-.700	.782	1.00	-.793	.785
timspk	-.009	-.069	.092	-.177	.275	.190	.598	.697	-.601	-.793	1.00	-.853
menps	-.025	.080	.001	.093	-.191	-.199	-.477	-.731	.373	.785	-.853	1.00

Table 6.7

Correlation Matrix for the Picnic Task

Correl	reform	falstar	replac	repetit	accura	Comp x	lofrun	spchrt	nofpas	totsiln	timspk	menps
reform	1.00	.825	.249	.487	-.017	-.157	-.179	.075	.038	.044	.161	-.055
falstar	.825	1.00	.470	.471	-.070	-.123	-.177	.066	.146	.115	.162	-.025
replac	.249	.470	1.00	.216	-.028	-.110	-.043	.120	.116	.100	-.006	.044
repetit	.487	.471	.216	1.00	-.101	-.185	-.299	-.070	.297	.302	-.275	.273
accura	-.017	-.070	-.028	-.101	1.00	.381	.536	.203	-.194	-.243	.271	-.300
compx	-.157	-.123	-.110	-.185	.381	1.00	.453	.485	-.235	-.293	.185	-.321
lofrun	-.179	-.177	-.043	-.299	.536	.453	1.00	.580	-.465	-.483	.571	-.512
spchrt	.075	.066	.120	-.070	.203	.485	.580	1.00	-.421	-.479	.430	-.544
nofpas	.038	.146	.116	.297	-.194	-.235	-.465	-.421	1.00	.929	-.730	.582
totsiln	.044	.115	.100	.302	-.243	-.293	-.483	-.479	.929	1.00	-.780	.771
timspk	.181	.162	-.006	-.275	.271	.185	.571	.430	-.730	-.780	1.00	-.831
menps	-.055	-.025	.044	.273	-.300	-.321	-.512	-.544	.582	.771	-.831	1.00

Table 6.8

Correlation Matrix for the Football Task

Correl	reform	falstar	replac	repetit	accura	Comp x	lofrun	spchrt	nofpas	totsiln	timspk	menps
reform	1.00	.868	.056	.380	.029	-.079	-.261	.045	.111	-.085	.186	-.179
falstar	.868	1.00	.333	.448	-.042	-.008	-.347	-.079	.235	.068	.064	-.034
replac	.056	.333	1.00	.173	-.172	.131	-.110	-.035	.207	.094	.001	.010
repetit	.380	.448	.173	1.00	-.194	-.196	-.417	-.223	.348	.277	-.181	.217
accura	.029	-.042	-.172	-.194	1.00	.413	.357	.267	-.254	-.249	.254	-.231
compx	-.079	-.008	.131	-.196	.413	1.00	.425	.364	-.179	-.210	.253	-.200
lofrun	-.261	-.347	-.110	-.416	.357	.425	1.00	.767	-.582	-.584	.609	-.530
spchrt	.045	-.079	-.035	-.223	.267	.364	.767	1.00	-.716	-.781	.749	-.701
nofpas	.111	.235	.207	.348	-.254	-.179	-.582	-.716	1.00	.799	-.735	.472
totsiln	-.085	.068	.094	.277	-.249	-.210	-.584	-.781	.799	1.00	-.915	.864
timspk	.186	.064	.001	-.181	.254	.253	.609	.749	-.735	-.915	1.00	-.845
menps	-.179	-.034	.010	.217	-.231	-.200	-.530	-.701	.742	.864	-.845	1.00

6.2.2 MANOVA: Effects of the Independent Variables

In order to investigate the effect of task structure, planning condition and proficiency level on language performance, a repeated measures MANOVA was carried out. In effect, running a repeated measures MANOVA was necessary to indicate whether there were any differences among the performances as a result of the effect of the independent variables, i.e. task structure, planning conditions and proficiency levels. As running MANOVA with all the dependent variables, i.e. the 12 measures of fluency, accuracy and complexity, would make the results less clear (Tabachnic & Fidell, 1996), it was necessary to select representatives of each group of measures. Based on the results of the factor analyses, four dependent variables were selected: number of false starts, number of pauses, accuracy and complexity. The criterion for selecting one measure from the temporal and one from the repair fluency in each factor group was the consistency in loadings of these measures across the four tasks. Although all measures of temporal fluency consistently loaded on Factor 1, total silence was selected because it had a consistently high loading of between .87 and .95 across the four tasks. For a similar reason, false start was selected from among other measures of repair fluency since it had a consistently high loading of between .90 and .94 on Factor 2. As regards language form, both measures of complexity and

accuracy were the dependent variables of the study and therefore had to be included in the analysis. The independent variables of the analysis were planning and proficiency level, each with two levels, and task structure with four levels. The results from the repeated measures MANOVA are presented in Table 6.9.

Table6.9

Results of Repeated Measures MANOVA

Between-Participants Effect

Effects	**Pillai's Value**	***F***	***BGdf***	***WGdf***	**Sig.**
Planning	.179	4.00	4	73	.006*
Proficiency	.374	10.89	4	73	.001*
Planning x Proficiency	.103	2.09	4	73	.09

Within-Participants Effects

Effects	**Pillai's Value**	***F***	***BGdf***	***WGdf***	**Sig.**
Task	.754	16.78	12	65	.001*
Task x Planning	.263	1.93	12	65	.16
Task x Proficiency	.278	2.09	12	65	.12
Task x Planning x proficiency	.288	2.19	12	65	.09

* Significant differences are reached.

With regard to the between-participants effect, the analysis revealed a significant difference between the planners and non-planners (Pillais = .179, $F = 4.00$, $P = .006$) and between low and high proficiency level (Pillais = .374, $F = 10.89$, $P = .001$). A significant difference was further observed across the four tasks as the within-participants variable (Pillais = .754, $F = 16.78$, $P = .001$) with differences being concentrated on the number of pauses (Walkman: M = 26, Unlucky Man: M = 22, Picnic: M = 19, Football: M = 18); on complexity (Walkman: M = 1.36, Unlucky Man: M = 1.32, Picnic: M = 1.60, Football: M = 1.43); on false starts (Walkman: M = 5.2, Unlucky Man: M = 4.41, Picnic: M = 4.59, Football: M = 3.97); and on

accuracy (Walkman: M = .30, Unlucky Man: M = .30, Picnic: M = .43, Football: M = .42). When the results for the dependent variables were considered separately through Univariate F test, using a Bonferoni adjusted alpha level[1] (recommended by Tabachnic and Fidell, 1996), significance was reached for all four measures as a result of task effect. However, the only significant result in the interaction effects between task and proficiency level seen was for complexity. Results of the Univariate F test are demonstrated in Table 6.10.

Table 6.10
Univariate Test of Within-Participant Effect

Source	Measure	Sum of squares	*Df*	Mean square	*F*	Sig.
Task	No. of pauses	3047.55	3	1015.85	20.21	.001*
	Complexity	3.53	3	1.18	25.65	.001*
	False start	63.00	3	21.00	3.95	.009*
	Accuracy	1.22	3	.407	29.80	.001*
Task x Planning	No. of pauses	64.55	3	21.52	.42	.73
	Complexity	.431	3	.144	3.126	.02
	False start	24.85	3	8.28	1.55	.2
	Accuracy	.082	3	.027	2.015	.11
Task x Prof.	No. of pauses	6.83	3	2.27	.045	.98
	Complexity	.465	3	.155	3.37	.01*
	False start	21.60	3	7.20	1.35	.25
	Accuracy	.123	3	.041	3.00	.03
Task x Pl. x Prof.	No. of pauses	291.9	3	97.3	1.93	.124
	Complexity	.184	3	.06	1.33	.26
	False start	29.18	3	9.72	1.83	.14
	Accuracy	.034	3	.011	.83	.36

* Significant differences are reached.

[1]A Bonferoni adjustment to alpha level is usually adopted to prevent an inflated risk of Type 1 errors. It is, in fact, a more stringent level to avoid rejecting a null hypothesis when it is true.

So far the results indicate that there is a statistically significant difference across the tasks. The results of the pairwise comparisons of all the four measures will show where the specific statistical differences among the tasks are located. Regarding the number of pauses, the two structured tasks are not different from one another but are significantly better in fluency than the two unstructured tasks. Similarly, accuracy measures of the four tasks show that the two unstructured tasks are not significantly different from each other but are different from both structured tasks. In other words, performance in the Football and Picnic tasks was more accurate and with fewer pauses than performance on Walkman and Unlucky Man. Regarding complexity, the unstructured tasks were not significantly different from one another nor from the Football task but the three tasks are significantly different from the Picnic task. Performance in the Picnic task, in effect, was significantly more complex (Mean of complexity measure for Picnic = 1.60, SD = .03) than performance in the other three tasks. Nevertheless, the Football task is the second most complex task (M = 1.43, SD = .03), Walkman (M = 1.36, SD = .02) is the third and Unlucky Man (M = 1.32, SD = .02) has elicited the least complex performance. As with the false starts, once more the unstructured tasks are not different from one another nor from the Picnic task. Performance in the Football task, in fact, has the fewest number of false starts (M = 3.97, SD = .34); i.e. it is the most fluent performance. The details of the pairwise comparisons across the four tasks are given in Tables 6.11 to 6.14.

Table 6.11

Pairwise Comparison between Tasks: No. of Pauses

Tasks	Walkman	Unlucky	Picnic	Football
Walkman	-	NS	.001	.001
Unlucky		-	.001	.001
Picnic			-	NS
Football				-

Table 6.12

Pairwise Comparison between Tasks: Complexity

Tasks	Walkman	Unlucky	Picnic	Football
Walkman	-	NS	.001	NS
Unlucky		-	.001	.001
Picnic			-	.001
Football				-

Table 6.13

Pairwise Comparison between Tasks: False Start

Tasks	Walkman	Unlucky	Picnic	Football
Walkman	-	NS	NS	.001
Unlucky		-	NS	NS
Picnic			-	NS
Football				-

Table 6.14

Pairwise Comparison between Tasks: Accuracy

Tasks	Walkman	Unlucky	Picnic	Football
Walkman	-	NS	.001	.001
Unlucky		-	.001	.001
Picnic			-	NS
Football				-

6.2.3 ANOVA: Effects of Task Structure

The next statistical analysis employed in this study deals with the effects of task structure on the individual dependent variables of the study. In order to determine whether there were any significant differences among the performances on different tasks, a repeated-measures ANOVA was conducted for each dependent variable, i.e. measures of fluency, accuracy and complexity. Where significance was reached a Scheffe test was conducted to establish where the differences were located. In case of non-significant results, pairwise comparisons were run to explore the differences between pairs of the tasks. The results of the comparisons are given in Table 6.15, with the *F*-values, the significance level, Eta squared, means of the measure for each task, standard deviations, and an indication of where the differences reached significance.

Table 6.15

Results of ANOVA: Effects of Task Structure

Measures	*F*	*P*	*Eta Sq.*	Task Structure				*Sig. Multiple Comparison*	*Sig. Pairwise Comparison*
				Walkman	**Unlucky Man**	**Picnic**	**Football**		
Total silence	3.80	.04*	.156	29.62 (SD = 23.87)	27.39 (SD = 25.09)	20.57 (SD = 22.72)	19.47 (SD = 19.56)	F vs. W	F P vs. U W
Length of run	4.99	.008*	.151	3.59 (SD = 1.07)	3.29 (SD = 1.09)	3.85 (SD = 1.41)	4.05 (SD = 1.57)	F P vs. U	F vs. U W
Pause length	2.72	.16	.069	1.02 (SD = .54)	1.09 (SD = .61)	.9 (SD = .38)	.9 (SD = .4)	-----	F vs. U
No. of pauses	6.90	.001*	.205	26.6 (SD = 12.71)	22.51 (SD = 11.46)	19.92 (SD = 12.41)	18.47 (SD = 11.89)	F P vs. W	F P vs. U W
Prop. time spoken	6.45	.001*	.231	.71 (SD = .15)	.69 (SD = .15)	.76 (SD = .13)	.79 (SD = .13)	F P vs. U	F P vs. U W
Speech rate	1.57	.19	.072	94.66 (SD = 29.42)	87.76 (SD = 29.95)	99.27 (SD = 41.77)	94.85 (SD = 33.09)	-----	-----

False start	1.87	.13	.40	5.2 (SD = 3.88)	4.41 (SD = 3.35)	4.58 (SD = 3.05)	3.96 (SD = 3.02)	-----	F vs. W
Reformulation	2.79	.16	.55	3.06 (SD = 2.4)	2.28 (SD = 1.96)	2.79 (SD = 2.09)	2.21 (SD = 2.17)	-----	F vs. W U vs. W
Replacement	1.29	.27	.019	.61 (SD = .83)	.71 (SD = 1.41)	.43 (SD = .69)	.48 (SD = .79)	-----	-----
Repetition	.53	.65	.008	4.28 (SD = 4.58)	3.88 (SD = 4.69)	3.46 (SD = 3.45)	3.78 (SD = 3.70)	-----	-----
Accuracy	9.79	.001*	.267	.30 (SD = .20)	.30 (SD = .18)	.43 (SD = .19)	.42 (SD = .22)	F P vs. U W	-----
Complexity	15.19	.001*	.234	1.36 (SD = .28)	1.31 (SD = .20)	1.59 (SD = .33)	1.43 (SD = .28)	P vs. F U W	P vs. U W

* Significant differences are reached across tasks.

F = Football, P = Picnic, U = Unlucky Man, W = Walkman

6.3 Results

6.3.1 Hypotheses 1 and 2

Hypothesis 1a predicted that language performance on structured tasks would be more fluent than performance on unstructured tasks. Fluency was measured by total amount of silence, number of pauses, mean length of pauses, length of run, proportion of time spoken, and speech rate as different measures of temporal fluency and false start, reformulation, replacement and repetition as measures of repair fluency. The results of the ANOVAs show that differences across the four tasks were clearly significant on the measures of total amount of silence, length of run, proportion of time spoken, number of pauses and false start (See Table 6.15). For all these measures the differences reached significance with one or both of the structured tasks being more fluent than one or both unstructured tasks. For the number of pauses and proportion of time spoken the two structured tasks, i.e. Football and Picnic, were significantly more fluent than the two unstructured tasks, i.e. Unlucky Man and Walkman. For the length of run Football was significantly better in fluency than Unlucky Man; for the total amount of silence Football and Picnic were different from Walkman and Unlucky Man, and for false start Football was significantly more fluent than Walkman. Hence, it can be concluded that Hypothesis 1a is broadly confirmed since performance in the structured tasks was clearly more fluent than performance in unstructured tasks.

Before discussing Hypothesis 1b, let us consider Hypothesis 2, which is directly related to the effect of task structure. Hypothesis 2 predicted that the effect of task structure on fluency would be progressively greater in line with the degree of structure defined in the study. In other words, it was hypothesized that performance on the Football task is more fluent than performance on Picnic, which are respectively more

fluent than performances on Unlucky Man and Walkman. Mean scores for total amount of silence, number of pauses and speaking time confirm this hypothesis as they have progressively increased with the degree of structure. In fact, Walkman has elicited the least fluent performance and Football has elicited the most fluent performance of the four tasks with Picnic very close to Football and Unlucky Man very close to Walkman. Table 6.16 shows the mean scores of these three measures on different tasks.

Table 6.16

Mean Scores of Fluency across Tasks

Measures	Football	Picnic	Unlucky Man	Walkman
No. of pauses	18	19	22	26
Total silence	19	20	27	29
Prop. of time spoken	.79	.76	.69	.71

Scores on other measures of fluency also support this hypothesis to a large extent. For pause length, Football and Picnic have elicited the most fluent performances, whereas Walkman has elicited the least fluent performance. With false starts there is a clear progression: performance in Football is the most and performance in Walkman is the least fluent performance. But the progression is not exactly as it has been predicted for Picnic and Unlucky Man. For the length of run, speaking time, and reformulation, Football has elicited the most and Unlucky Man the least fluent performance. Scores on repetition show that the Picnic task has elicited the most repetitive performance and Football has elicited the least amount of repetition. With regard to replacements, Picnic has elicited the most and Unlucky Man the least fluent performances.

Hypothesis 1b predicted that performance on structured tasks would generate more accurate language. The results of the ANOVA reveal that a significant difference was

reached between the structured tasks and unstructured tasks with regard to accuracy measure ($F = 9.79$, $P < .001$, $\eta^2 = .267$). The results of Scheffe test of post hoc comparison further showed that the two structured tasks generated significantly more accurate language than the two unstructured tasks. The location of the differences across the tasks for accuracy is shown in Table 6.17. The multiple comparisons revealed that the structured tasks, i.e. Football and Picnic, are not different from each other but are significantly different from the unstructured tasks, i.e. Walkman and Unlucky Man.

Table 6.17

Multiple Comparison between Tasks: Accuracy

Tasks	Walkman	Unlucky	Picnic	Football
Walkman	-	NS	.001	.001
Unlucky		-	.001	.001
Picnic			-	NS
Football				-

Hypothesis 2 further predicted that language performance in structured tasks would be, as a function of degree of structure as defined in Chapters IV and V, progressively more accurate than the performance on unstructured tasks. The mean scores of the four tasks on the accuracy measure reveal that there is a relatively high degree of progression in accuracy in line with the predicted degree of structure. In other words, the accuracy indices are statistically higher on the structured tasks. However, within the groups of structured or unstructured tasks, the figures are very close to one another. Picnic, with a mean accuracy of .43, has elicited the most accurate performance, Football with a mean accuracy of .42 comes second, Unlucky Man and Walkman with a mean accuracy of .30 have elicited the least accurate performances. Hypothesis 2 thus receives partial but important confirmation from the accuracy measures.

Hypothesis 1c predicted that there would be no significant difference between the complexity of the performances elicited by structured tasks and that of unstructured tasks. However, the results of the ANOVA show that there is a significant difference between a structured task and the other three tasks. Picnic (M = 1.59) has elicited significantly greater complexity of language as compared with Football (M = 1.43), Unlucky Man (1.31), and Walkman (M = 1.36). However, it should be noted that Football has elicited the second most complex performance. In fact, the mean scores of the complexity across tasks imply that the performance in structured tasks tends to be more complex than the performance on unstructured tasks. Therefore, Hypothesis 1c can not be thoroughly confirmed. It is worth mentioning that the figures for the effect size for measures of accuracy, complexity, number of pauses and proportion of time spoken are noticeable, suggesting that a great amount of the total variance in these measures is explained by the independent variable, i.e. task structure (Tabachnic & Fidell, 1996).

6.3.2 Hypothesis 3

Hypothesis 3 predicted that language performance under planned conditions would be more fluent, more accurate and more complex than that produced under unplanned conditions. In order to locate the effect of planning time on different measures of fluency, accuracy and complexity, a series of t-tests were carried out on each dependent variable. Furthermore, to compare the effect of pre-task planning with the effect of language proficiency on the dependent variables, a number of t-tests were performed for each measure of fluency, accuracy and complexity for the two levels of proficiency. The results of the t-tests for planning conditions are presented in Table 6.18a and for proficiency levels in Table 6.18b.

Table 6.18a

Results of T-Tests: Effects of Planning Conditions

Measures	*T*	*P*	Unplanned	Planned
Total silence	4.16	.001*	29.53 (SD = 27.42)	19 (SD = 16.55)
Length of run	4.16	.001*	3.39 (SD = 1.18)	4.00 (SD = 1.4)
Pause length	5.93	.001*	1.14 (SD = .6)	.81 (SD = .3)
No. of pauses	1.68	.18	23.05 (SD = 12.91)	20.70 (SD = 11.92)
Prop. time spoken	5.80	.001*	.69 (SD = .15)	.78 (SD = .13)
Speech rate	3.14	.008*	88.23 (SD = 38.04)	100.04 (SD = 28.36)
False start	.21	.82	4.5 (SD = 3.17)	4.58 (SD = 3.55)
Reformulation	1.12	.26	2.45 (SD = 2.03)	2.72 (SD = 2.33)
Replacement	.57	.56	.53 (SD = 1.12)	.59 (SD = .81)
Repetition	.78	.43	3.67 (SD = 3.57)	4.00 (SD = 4.63)
Accuracy	5.52	.001*	.30 (SD = .19)	.42 (SD = .21)
Complexity	2.23	.04*	1.38 (SD = .29)	1.46 (SD = .29)

* Significant differences are reached.

Table 6.18b

Results of T-Tests: Effects of Language Proficiency

Measures	*T*	*P*	Elementary	Intermediate
Total silence	3.07	.004*	28.22 (SD = 22.94)	20.32 (SD = 22.88)
Length of run	6.12	.001*	3.26 (SD = 1.10)	4.12 (SD = 1.46)
Pause length	3.72	.001*	1.08 (SD = .56)	.87 (SD = .45)
No. of pauses	2.00	.08*	23.26 (SD = 11.28)	20.48 (SD = 13.42)
Prop. time spoken	2.65	.01*	.71 (SD = .15)	.76 (SD = .14)
Speech rate	6.72	.001*	82.15 (SD = 25.82)	106.12 (SD = 36.94)
False start	2.87	.008*	5.07 (SD = 3.63)	4.00 (SD = 3.00)
Reformulation	2.46	.03*	2.88 (SD = 2.53)	2.28 (SD = 1.74)
Replacement	1.95	.1	.66 (SD = 1.18)	.45 (SD = .70)
Repetition	1.35	.34	4.16 (SD = 3.82)	3.54 (SD = 4.41)
Accuracy	7.43	.001*	.28 (SD = .19)	.44 (SD = .20)
Complexity	6.62	.001*	1.32 (SD = .23)	1.53 (SD = .31)

* Significant differences are reached.

The results of the t-tests show that the effect of pre-task planning reached statistical significance for measures of total silence ($t = 4.16$, $P = .001$), length of run ($t = 4.16$, $P = .001$), pause length ($t = 5.93$, $P = .001$), proportion of time spoken ($t = 5.80$, $P = .001$) and speech rate ($t = 3.14$, $P = .002$). The mean scores for each measure show that performances were more fluent under planned conditions. With 5 fluency measures revealing significant difference under planned conditions, Hypothesis 3a can be broadly confirmed. Although measures of reformulations and number of pauses do not reach statistical significance, it can be clearly seen that performance under planned conditions tends to have fewer reformulations and pauses than unplanned performance. However, with other measures of repair fluency, a significant difference is not reached.

All measures of temporal fluency are significantly higher in the intermediate language proficiency group. False starts, reformulations and replacements are also significantly lower, indicating greater fluency, at the intermediate level. This shows that the language performance of high proficiency participants is more fluent than the performance of low proficiency participants. Interestingly, the effect of planning condition on the total amount of silence, pause length and proportion of time spoken is greater than the effect of language proficiency. It can be concluded that having the opportunity to plan would more effectively help participants to produce more fluent language than their being at a higher level of proficiency.

Hypothesis 3b predicted that language performance would be more accurate under planned conditions. Results of the t-tests show that accuracy has significantly improved under planned conditions ($t = 5.52$, $P = .001$). Therefore, this hypothesis receives clear and strong support from the accuracy measure. As expected, language performed by high proficiency participants is also significantly more accurate than

that produced by low proficiency participants ($t = 7.34$, $P = .001$). It should be noted that the effect of level of language proficiency on accuracy is greater than the effect of pre-task planning.

Hypothesis 3c predicted that language performance would be more complex under planned conditions. As can be seen in Table 6.18a, a significant difference is reached for the complexity of performance between the two planning conditions ($t = 2.32$, $P = .02$) with the planned group achieving a higher degree of complexity in their performance. Therefore, Hypothesis 3c can be confidently confirmed. The results of the t-tests also reveal that the effect of proficiency level on complexity seems to be greater than the effect of pre-task planning ($t = 6.62$, $P = .001$).

6.3.3 Hypotheses 4 and 5

Hypotheses 4 and 5 include predictions regarding the interactions among task structure, planning conditions and proficiency levels. In order to investigate the effects of these three independent variables and to find out whether task structure has any interaction with planning condition and/or proficiency level a series of three-way ANOVAs were carried out. When significance was reached, Scheffe tests of post-hoc comparison were run on different levels of task structure to establish where the significant differences were located. The three-way ANOVAs were carried out on the four main measures of total amount of silence, false starts, accuracy and complexity. As discussed earlier, the results of the factor analyses revealed that all measures of temporal fluency highly loaded on Factor 1 and all measures of repair fluency loaded on Factor 2. Based on the results of factor analyses, it was decided that for some further analyses, measures of total amount of silence and false starts would be selected to represent temporal fluency and repair fluency respectively. Measures of

accuracy and complexity are themselves two dependent variables of the study and would represent the two aspects of form (See section 6.2.1 for a detailed discussion and results of the factor analyses). Table 6.19 shows the results of the three-way ANOVA for total amount of silence.

Hypothesis 4 predicted that the effect of planning would be, as a function of degree of structure defined in the current study, progressively more effective for the structured tasks with respect to fluency, accuracy and complexity. To be able to say whether this Hypothesis is confirmed, it is necessary to have a detailed comparison of the gain scores and percentage of changes that have resulted from the effects of pre-task planning for each of the tasks. Table 6.20 shows the mean scores, gain score and percentage of change in the total amount of silence that planners and non-planners had across tasks.

Table 6.19

Results of Three-way ANOVA: Total Silence

Source	Type III Sum of Squares	*Df*	Mean Square	*F*	*P*	Eta. Square
Planning	8876.68	1	8876.68	17.96	.001*	.056
Proficiency Level	4978.03	1	4978.03	10.07	.004*	.032
Task	6000.578	3	2000.19	4.04	.03*	.038
Planning x Task	71.26	3	23.75	.04	.986	.000
Planning x Prof.	944.86	1	944.86	1.91	.168	.006
Prof. Level x Task	483.89	3	161.29	.32	.806	.003
Plan x Prof. x Task	393.23	3	131.08	.26	.850	.003

* Significant differences are reached.

Table 6.20

Mean Scores for Total Silence across Tasks

	Football	Picnic	Unlucky	Walkman
Unplanned	25.45	25.30	32.78	34.57
Planned	13.49	15.83	22.01	24.66
Gain scores	11.96	9.47	10.68	9.91
% Change	46%	37%	33%	28%

As the comparison of the mean scores on the fluency measure of total silence reveals, participants have benefited from planning time across all the tasks. Furthermore, having been given the planning time, participants' performance on the structured tasks improved more than their performance on unstructured tasks. The percentages of change clearly show that planners were more fluent on the structured tasks as the figures for Football (46%) and Picnic (37%) are higher than the figures for Unlucky Man (33%) and Walkman (28%). Therefore there is a clear progression in fluency, in line with the degree of structure, observed through the percentage of change across tasks. Hence, Hypothesis 4 can be confirmed as far as temporal fluency is concerned.

Before dealing with other measures for Hypothesis 4, let us consider Hypothesis 5a which predicted that, regarding the fluency of performance, high proficiency participants would benefit more from the planning time. In fact, this Hypothesis predicts that the performance of the planners at high-proficiency level would improve more than the performance of the planners at a low-proficiency level in terms of fluency of their performance. Table 6.21 shows the mean scores, gain scores and the percentage of change for the planners and non-planners at both levels of proficiency.

Table 6.21

Mean Scores for Total Silence across Planning Conditions and Proficiency Levels

	Low Proficiency	High Proficiency	% Change in Proficiency
Unplanned	35.19	23.87	32%
Planned	21.22	16.77	21%
Gain Score	13.98	7.1	-
% Change	39%	30%	-

A comparison between the mean scores for total silence of the low-proficiency and high-proficiency participants under planned and unplanned conditions show that the high-proficiency participants (with 30% total silence) have benefited from the planning time less than the low-proficiency participants did(with 39% total silence). In other words, low-proficiency participants, when having been given the planning time, were able to reduce their total amount of silence more effectively. Therefore, Hypothesis 5a can not be confirmed for the temporal measure of fluency.

False starts are one of the four repair fluency measures employed in the current study, which is, as discussed earlier, selected to represent repair fluency. A three-way ANOVA was carried out to find the effects of the independent variables, i.e. task structure, planning condition and proficiency level, as well as any interaction among the independent variables. Results of the three-way ANOVA for false starts are represented in Table 6.22.

Table 6.22

Results of Three-way ANOVA: False Start

Source	Type III Sum of Squares	*Df*	Mean Square	*F*	*P*	Eta. Square
Planning	.52	1	.52	.04	.82	.000
Proficiency Level	91.378	1	91.37	8.32	.004*	.027
Task	63.00	3	21.00	1.91	.12	.019
Planning x Task	24.85	3	8.26	.75	.52	.007
Planning x Prof.	45.75	1	45.75	4.17	.04*	.014
Prof. Level x Task	21.60	3	7.20	.65	.57	.006
Plan x Prof. x Task	29.18	3	9.72	.88	.44	.009

* Significant differences are reached.

As mentioned earlier, Hypothesis 4 predicted that the effect of planning would be, as a function of degree of structure progressively more effective for the structured tasks with respect to fluency, accuracy and complexity. Table 6.23 shows the mean scores, gain score and percentage of change in the number of false starts planners and non-planners have made across tasks.

Table 6.23

Mean Scores for False Starts across Tasks

	Football	Picnic	Unlucky	Walkman
Unplanned	3.80	4.32	4.85	5.02
Planned	4.12	4.85	3.97	5.37
Gain scores	-.32	-.53	.88	-.35
% Change	-8%	-12%	18%	-6%

The mean scores of the false starts under the unplanned condition do not show a clear progression, in line with the degree of structure, in fluency across the tasks as predicted in Hypothesis 4. As can be seen in Table 6.23, the percentage of change for Football, Picnic and Walkman are negative, indicating that these three tasks had more false starts under planned conditions, whereas Unlucky Man had a positive change of 18%. This means under planned conditions, performances on Football, Picnic and Walkman are less fluent in terms of the number of false starts. There is not a clear picture of any consistent changes across the tasks with regard to false starts. In effect, when planning time is provided, performance on Unlucky Man is the only one that has benefited from planning in terms of the number of false starts. Performance on the other three tasks contains more false starts under planned conditions. As a result, Hypothesis 4 does not receive confirmation regarding the measure of false start.

Hypothesis 5a also predicted that high-proficiency participants would generally perform better and particularly benefit more from pre-task planning, than low-proficiency participants with regard to fluency measures. Table 6.24 presents the

mean scores, gain scores and percentage of change in fluency measure of false start for low and high proficiency-level participants and under both planned and unplanned conditions.

Table 6.24

Mean Scores for False Starts across Planning Conditions and Proficiency Levels

	Low Proficiency	High Proficiency	% Change in Proficiency
Unplanned	5.41	3.58	33%
Planned	4.73	4.42	6%
Gain Score	.68	-.84	
% Change	12%	-23%	

As the results indicate, high proficiency candidates have generally performed better than low-proficiency candidates under both planned and unplanned conditions, with 33% and 6% of change under unplanned and planned conditions respectively. However, percentages of change resulted from planning conditions reveal that low-proficiency participants have benefited from planning time (%12) more than high-proficiency participants (-23%) have. Therefore, Hypothesis 5a receives partial support from the measure of false start.

A three-way ANOVA was carried out to investigate the effect of the independent variables on the accuracy measure as well as any interaction among the independent variables. Accuracy was measured by the ratio of error-free clauses to total number of clauses. The results of the three-way ANOVA on the accuracy measure, presented in Table 6.25 below, reveal that planners' performances were significantly different from those of non-planners ($F = 44.55, P = .001, \eta^2 = .129$). Performances are also significantly different at the two proficiency levels ($F = 64.82, P = .001, \eta^2 = .178$) and across the four tasks ($F = 13.63, P = .001, \eta^2 = .155$). The interaction between the planning condition and proficiency level is also significant ($F = 12.28, P = .001, \eta^2 = .03$).

Table 6.25

Results of Three-way ANOVA: Accuracy

Source	Type III Sum of Squares	*Df*	Mean Square	*F*	*P*	Eta. Square
Planning	1.33	1	1.33	44.55	.001*	.129
Proficiency Level	1.94	1	1.94	64.82	.001*	.178
Task	1.22	3	1.22	13.63	.001*	.155
Planning x Task	.078	3	.026	.86	.45	.009
Planning x Prof.	.36	1	.36	12.28	.001*	.03
Prof. Level x Task	.116	3	.038	.27	.27	.01
Plan x Prof. x Task	.03	3	.01	.44	.723	.004

* Significant differences are reached.

Hypothesis 4, further predicted that the effect of planning would be, as a function of degree of structure defined in chapter three, progressively greater for the structured tasks in terms of the accuracy measure. Table 6.26 below shows the mean scores, gain scores and percentage of change for the accuracy measure across tasks.

Table 6.26

Mean Scores for Accuracy across Tasks

	Football	Picnic	Unlucky	Walkman
Unplanned	.38	.35	.22	.25
Planned	.47	.52	.38	.37
Gain scores	.09	.17	.16	.12
% Change	23%	48%	72%	48%

The results of the percentage of change in the accuracy measure across the tasks reveal that planning time has been more effective with the unstructured tasks. The amount of change in accuracy which has resulted from planning time is 23% for Football and 48% for Picnic, whereas the unstructured tasks show a greater amount of

change in accuracy under planned conditions with 72% of change for Unlucky Man and 48% for Walkman. It should be noted that performance on Walkman and Picnic has equally benefited from pre-task planning in terms of the accuracy measure. Therefore, Hypothesis 4 does not receive broad support from the accuracy measure.

Hypothesis 5b predicted that high-proficiency participants would generally perform better and particularly benefit more from planning time, than the low-proficiency participants in terms of the accuracy of their performance. The results of the three-way ANOVA on the accuracy measure showed that there was a significant difference between the high-proficiency and low-proficiency participants in terms of the accuracy of their performance. However, detailed comparisons are needed to explore how accurate participants have been at different proficiency levels and under different planning conditions. Table 6.27 demonstrates mean scores, gain scores and percentage of change for accuracy across planning conditions and proficiency levels.

Table 6.27

Mean Scores for Accuracy across Planning Conditions and Proficiency Levels

	Low Proficiency	High Proficiency	% Change in Proficiency
Unplanned	.26	.35	34%
Planned	.32	.55	71%
Gain Score	.06	.2	
% Change	23%	57%	

Comparing mean scores and percentage of change for the accuracy measure across the planning conditions and proficiency levels reveals that the change in the accuracy measure, resulting from the level of language proficiency, is greater for the planned group (71%) than the unplanned group (34%). Furthermore, results show that the effect of planning on accuracy is greater for high-proficiency participants, with 57% more accuracy, as compared to low-proficiency participants who achieved 23% more

accuracy. As a result, it can be concluded that Hypothesis 5b receives confirmation from the accuracy measure.

The last three-way ANOVA was carried out to investigate the effects of the independent variables, i.e. task structure, planning condition and proficiency level, on the complexity measure and to see whether there are any interactions among the independent variables. Results of the three-way ANOVA on the complexity measure are given in Table 6.28.

Table 6.28

Results of Three-way ANOVA: Complexity

Source	Type III Sum of Squares	*Df*	Mean Square	*F*	*P*	Eta. Squared
Planning	.467	1	.467	7.27	.01*	.023
Proficiency Level	3.40	1	3.41	52.93	.001*	.148
Task	3.53	3	1.18	18.36	.001*	.153
Planning x Task	.431	3	.144	2.23	.16	.022
Planning x Prof.	.047	1	.047	.744	.389	.002
Prof. Level x Task	.465	3	.155	.2.41	.124	.023
Plan x Prof. x Task	.184	3	.061	.952	.416	.009

* Significant differences are reached.

The results of the three-way ANOVA on the complexity measure, presented in Table 6.28, reveal that planners' performances were significantly different from those of non-planners ($F = 7.27$, $P = .01$, $\eta^2 = .023$). Performances are also significantly different at the two proficiency levels ($F = 52.93$, $P = .001$, $\eta^2 = .148$) and across the four tasks ($F = 18.36$, $P = .001$, $\eta^2 = .153$). However, no interaction between planning conditions and proficiency level reached a statistically significant level. It should be noted that the effect size for planning is a small one but the effect size (Eta

Squared) for task is interestingly even greater than the effect size for proficiency level.

Hypothesis 4 predicted that, regarding the complexity measure, the effect of planning would not be, as a function of degree of structure, progressively greater for the structured tasks. In other words, although planners are hypothesized to produce more complex language, this amount of complexity would not necessarily increase with the degree of structure across the tasks. Table 6.29 shows mean scores, gain scores and percentage of change for the complexity measure across the four tasks and the planning conditions.

Table 6.29

Mean Scores for Complexity across Tasks and Planning conditions

	Football	Picnic	Unlucky	Walkman
Unplanned	1.41	1.57	1.30	1.26
Planned	1.44	1.62	1.33	1.46
Gain scores	.03	.05	.03	.20
% Change	2%	3%	2%	15%

The mean scores and percentage of change across the tasks suggest that there is no clear progression in the complexity measures across the tasks when planning time is available. Some of the percentages of change in this measure, e.g. 2% or 3%, are really small and would not be considered as noticeable differences. The results also indicate that pre-task planning does not have a clear progressive effect on the complexity of tasks, as a function of degree of structure. Hypothesis 4, therefore, receives general confirmation from the complexity measure. However, as the mean scores and percentage of change for total silence, false starts and accuracy measures have partially confirmed this Hypothesis, it can be claimed that Hypothesis 4 is only partially confirmed.

Regarding complexity of performance, Hypothesis 5c predicted that high proficiency participants would generally produce more complex language and particularly benefit

more from planning, than low-proficiency participants. Table 6.30 shows the mean scores, gain scores and percentage of change in complexity measures across tasks, planning conditions and proficiency levels.

The mean scores of the complexity measure generally show that the performance of high-proficiency participants was more complex than the performance of low-proficiency participants. The percentage of change across the planning conditions also reveal that high-proficiency participants under the planned conditions produce 17% more complex language as compared to14% more complex language produced under the unplanned conditions. The percentages of change also reveal that high proficiency participants benefited more from planning time in producing more complex language, as 4% of change is seen for low-proficiency and 6% of change is witnessed for high-proficiency groups. The percentages of change scores suggest that Hypothesis 5c receives general confirmation.

Table 6.30

Mean Scores for Complexity across Planning Conditions and Proficiency Levels

	Low Proficiency	High Proficiency	% Change in Proficiency
Unplanned	1.29	1.48	14%
Planned	1.35	1.58	17%
Gain Score	.06	.10	
% Change	4%	6%	

6.3.4 Hypothesis 6

Hypothesis 6 predicted that high-proficiency participants would benefit more from pre-task planning while performing the unstructured tasks. In other words, it was hypothesized that the performance of high-proficiency participants would be more fluent, accurate and complex on unstructured tasks when they had time to plan. Therefore, it is necessary to compare the performance of the high-proficiency participants for measures of fluency, accuracy and complexity across the four tasks.

Tables 6.31 shows the mean scores, gain scores and percentage of change for total silence of high-proficiency participants across the tasks and the two planning conditions.

Table 6.31

Percentage of Change for Total Silence of High-Proficiency Level

	Football	Picnic	Unlucky Man	Walkman
Unplanned	17.38	22.04	27.83	28.23
Planned	10.69	12.44	22.42	21.54
Gain scores	6.69	9.6	5.41	6.69
% Change	38%	43%	19%	23%

Table 6.31 shows that the percentages of change in total silence, resulting from the planning condition, suggest that high-proficiency participants produced more fluent language on structured tasks when they were given pre-task planning opportunity. Football with 38% and Picnic with 43% of change in their total amount of silence were more fluent than Unlucky Man with 19% and Walkman with 23% of change.

The same comparison is required to test whether high-proficiency participants were more fluent, in terms of the number of false starts, on structured tasks when they had time to plan.

Table 6.32

Percentage of Change for False Start of High-Proficiency Level

	Football	Picnic	Unlucky Man	Walkman
Unplanned	3.15	4.10	3.15	3.95
Planned	4.00	4.70	4.00	5.75
Gain scores	.85	.60	.85	1.80
% Change	27%	14%	27%	45%

Regarding the measure of false starts, the percentages of change which resulted from the planning conditions suggest that high-proficiency participants produced slightly more fluent language on structured tasks when they were given the planning time (It should be noted that higher figures of false starts indicates more frequent use of false

starts which in turn reflects less fluency in performances). The comparison between Picnic with only 14% increase and Walkman with 45% increase in false starts indicates that more fluency is associated with performances on the Picnic task. Nevertheless, the figures for Football and Walkman, with 27% of more false starts, suggest that there is no difference between these two tasks, in terms of the false start measure of fluency.

In order to find out whether Hypothesis 6 receives confirmation from the accuracy measure, a comparison of the performance of the high-proficiency participants across the four tasks is needed. Tables 6.33 shows the mean scores, gain scores and percentage of change for the accuracy of the performance of the high-proficiency participants across the tasks and the two planning conditions.

Table 6.33

Percentage of Change for Accuracy of High-Proficiency Level

	Football	Picnic	Unlucky Man	Walkman
Unplanned	.42	.39	.23	.33
Planned	.57	.61	.49	.50
Gain scores	.15	.22	.26	.17
% Change	35%	56%	113%	51%

The figures for the accuracy measure in Table 6.33 reveal that performance on unstructured tasks has generally benefited more from planning time. It suggests that high proficiency participants have improved their accuracy more on the unstructured tasks as compared with the structured tasks. The percentages of change for the structured tasks, i.e. Football (35%) and Picnic (56%), tends to be lower than the percentages of change for the unstructured tasks, i.e. Unlucky Man (113%) and Walkman (51%).

To investigate whether Hypothesis 6 receives confirmation from the complexity measure, the percentages of change that occurred in the complexity of performance of

high-proficiency participants as a result of pre-task planning should be compared. Tables 6.34 shows the mean scores, gain scores and percentage of change for the complexity of the performance of the high-proficiency participants across the tasks and the two planning conditions.

Table 6.34

Percentage of Change for Complexity of High-Proficiency Level

	Football	Picnic	Unlucky Man	Walkman
Unplanned	1.46	1.75	1.39	1.31
Planned	1.55	1.77	1.40	1.59
Gain scores	.09	.02	.01	.28
% Change	6%	1%	1%	21%

The percentages of change in the complexity measure indicated in Table 6.34, which resulted from the effects of planning conditions, suggest that high-proficiency participants produced more complex language on one of the unstructured tasks when they were given the planning time. As the figures show, the complexity of performance on Picnic and Unlucky Man with 1% increase has had the same amount of change when planning time is provided to the high-proficiency participants. Football, with 6% increase is slightly higher, but Walkman, with 21% increase demonstrates the greatest increase in terms of the complexity measure resulted from planning conditions.

As regards Hypothesis 6, the results are quite mixed with some measures suggesting positive change elicited by the structured tasks and some measures indicating positive change elicited by the unstructured tasks. Therefore, Hypothesis 6 can not be thoroughly confirmed or rejected.

6.3.5 Hypothesis 7

This Hypothesis deals with the test-takers' perceptions of task difficulty. It was hypothesized that the test-takers of this study would perceive unstructured tasks as

more difficult than structured tasks and this would be in line with the predicted difficulty of task, in terms of task structure. To test Hypothesis 7, a three-way ANOVA was carried out in which responses to the task difficulty questionnaire items were taken as the dependent variable and task structure, planning and proficiency levels were the independent variables. Considering the Bonferoni adjusted alpha level (Tabachnic and Fidell, 1996), the results of the three-way ANOVA show a significant difference for task structure ($F = 32.63$, $P = .001$, $\eta^2 = .244$) and also a significant difference for the planning conditions ($F = 6.11$, $P = .02$, $\eta^2 = .02$). However, no significance was reached for the proficiency levels or the interaction between the independent variables. The figure indicating the effect size for the task variable is noticeable, suggesting that a great amount of the variance in perceptions of task difficulty relate to task structure. Table 6.35 shows the results of the three-way ANOVA on the test-takers' perceptions of task difficulty.

Table 6.35

Results of Three-way ANOVA on Perceptions of Task Difficulty

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*P*	Eta. Square
Task	42.10	3	14.03	32.63	.001*	.244
Planning	2.62	1	2.62	6.11	.02*	.02
Proficiency Level	1.12	1	1.12	2.62	.106	.009
Planning x Task	.58	3	.19	.45	.71	.004
Task x Prof.	.93	3	.311	.72	.53	.007
Prof. X Planning	.37	1	.37	.87	.34	.003
Plan x Prof. X Task	.33	3	.11	.25	.85	.003

* Significant differences are reached.

As discussed in Chapter V, in the questionnaires a rating scale of 1 to 4 was considered to describe the difficulty level of the tasks. In this scale, 1 referred to *'very easy'* and 4 to *'very difficult'* tasks, with 2 and 3 representing *'easy'* and *'difficult'* respectively. Mean scores for the perceptions of task difficulty across tasks under the two planning conditions are shown in Table 6.36. The comparison showed that the participants rated the two unstructured tasks, i.e. the Unlucky Man and Walkman tasks, as more difficult than the two structured tasks under both planning conditions.

Table 6.36

Mean Scores of Perceptions of Task Difficulty

Tasks	Football	Picnic	Unlucky Man	Walkman
Unplanned	1.90	1.95	2.67	2.55
Planned	1.80	1.62	2.52	2.40

A Scheffe test of post-hoc comparison was then carried out to explore where the significant differences were located across the tasks. The multiple comparisons showed that the two structured tasks, i.e. the Football and Picnic, were not statistically different from one another but were statistically different from the two unstructured tasks. Table 6.37 demonstrates the multiple comparisons across the four tasks.

Table 6.37

Multiple Comparisons on Perceptions of Task Difficulty

Tasks	Walkman	Unlucky	Picnic	Football
Walkman	-	NS	.001	.001
Unlucky		-	.001	.001
Picnic			-	NS
Football				-

Questionnaires of the planned group included a section on the usefulness of the planning time for each of the tasks. A two-way ANOVA was carried out to investigate whether the participants of the two proficiency levels found planning time more useful for any of the tasks. The results of the analysis did not reveal any significant

differences across the tasks or the proficiency levels. Table 6.38 below shows the results of the two-way ANOVA.

Table 6.38

Results of Three-way ANOVA on Usefulness of Planning Time

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*P*	Eta. Square
Task	1.53	3	.51	.81	.48	.016
Proficiency	.94	1	.94	1.50	.22	.010
Task x Prof. Level	6.25	3	2.08	3.33	.06	.063

As explained in Chapter V, in the last part of the questionnaires the participants, in both planned and unplanned groups, were asked to comment on the tasks and the test. Since only 17 comments were received on this question for both conditions, it is difficult to come to a general conclusion. Among the comments received, a number of the participants have expressed their gratitude for having been given the opportunity to take part in the research or to take such a test. Others have said they would like to have more similar tests of speaking as they will help them prepare for their real tests. Under the unplanned conditions, a number of the participants have complained about not having been given pre-task planning time so that they can plan to speak more accurately.

Under the planned conditions, as explained before, participants were given an extra piece of paper and were advised to take notes about what they wanted to say. In total, 12 participants used this opportunity and took some notes. However, investigations of these notes did not show a clear pattern of note taking or any emphasis on a specific area of language.

In the chapter that follows, I will first draw upon the various statistical analyses presented in this chapter to summarize the results. Then, I will attempt to discuss how these findings relate to the previous findings of SLA and LT and what research questions are raised in the light of the results of this study.

CHAPTER VII

Observations: Findings of Study One

7.1 Overview

In Study One, I attempted to investigate the effects of three characteristics and conditions of oral narrative tasks on the performance of 80 Iranian second language learners of English in an assessment setting: the degree of structure of the tasks, pre-task planning and the language proficiency level of the participants. It should be mentioned that, in order to have an assessment setting during the data collection, every effort was made to make sure the setting is very similar to the assessment settings usually created in similar testing studies (e.g. Elder et al., 2002). I further tried to explore whether participant perceptions of task difficulty were in line with the predicted difficulty of tasks in terms of task structure. In the current chapter, I will first summarize the results of this study and will then discuss how they relate to previous findings from the literature particularly to the issues raised in earlier chapters.

The results of the data analyses show that oral narrative tasks that contain a clear structure, whether problem-solution or sequential organization, will elicit significantly more fluent and more accurate performances. The results further indicate that there is a progressive increase, in line with the degree of task structure, in fluency and accuracy of performances. However, a number of different patterns of progression are observed for various fluency measures. Regarding the effects of pre-task

planning, results from the post-hoc analyses show that under planned conditions the test-takers are able to produce language which is significantly more fluent, more accurate and more complex. In the light of the planning time provided to the test-takers, the improvement achieved by the low-proficiency group is greater than the improvement made by the high-proficiency group. The results regarding the effects of the interaction between pre-task planning and proficiency level on the performance of the test-takers across tasks are mixed, indicating that for some of the measures the high-proficiency group and for other measures the low-proficiency group have benefited more from the planning time. For instance, the low proficiency group has benefited more from the planning time on the measures of false starts and the total amount of silence. However, with the measures of accuracy and complexity, the high proficiency group has benefited more from the planning time. The results from the analysis of the questionnaires on perceptions of task difficulty strongly confirm that the test-takers perceive unstructured tasks as more difficult than unstructured tasks under both planned and unplanned conditions. In the sections that follow, I will discuss issues related to each of the variables of the study in detail.

7.2 Discussing Findings of Study One

7.2.1 Effects of Task Structure

In line with the theoretical background provided in earlier chapters, it was hypothesized that task structure would have an effect on the fluency and accuracy of performance on oral narrative tasks. The results confirm that the structure of a task strongly influences test-takers' language performance in terms of fluency, accuracy and complexity. Comparison of the performances across the four tasks shows that performance on structured tasks is significantly more fluent and more accurate than

that on unstructured tasks. Scores for complexity are also generally higher on structured tasks with a significant difference between the Picnic task and the other three tasks. Results regarding fluency and accuracy confirm the findings of Foster and Skehan (1999) and Wigglesworth (2001) who have reported that more fluency and accuracy is generated in the performance on structured tasks. It appears that presence of structure in a task reduces the cognitive load of the task on the test-takers and allows them to allocate their attentional resources to different aspects of form as well as meaning. It could be argued that the test-takers might find unstructured tasks more difficult to perform since they would have to allocate some of their time and attention to understanding the task itself. As explained before, in this study task structure was operationalized in terms of the number of pictures that could be removed in a picture story without the main theme of the story being changed. Given that in unstructured tasks there is not a clear time line or a macrostructure and the sequence of the events is arbitrary, it is likely that the participants employ at least part of their attention to understanding this lack of structure. Since there are fewer attentional resources available while performing the unstructured tasks, the test-takers would not be able to attend to all aspects of their performance equally well. In contrast, the clear macrostructure, the timeline and the fixed sequence of events in structured tasks help the test-takers understand the task better. It appears that the test-takers who are performing the structured tasks would not need to spend much attention understanding the structure of the tasks. This means, they would have more attentional resources available to focus on what meanings they want to express and what forms they would prefer to employ in to convey their intended meanings.

The effect of task structure on fluency is clearly seen for four of the fluency measures, i.e. total amount of silence, length of run, number of pauses and proportion of time

spoken. In effect, the results reveal that there are significant differences between the structured and unstructured tasks, with higher indices of fluency being generated by the structured tasks. For three other measures of fluency – mean length of pauses, false starts and reformulations, a statistical significance was reached between the most structured task and one of the unstructured tasks, i.e. between Football and Unlucky Man or Walkman. In case of some of the measures, however, the pattern of the results is not totally straightforward. For instance, for false start, a progressive trend is seen between Football and Walkman but the comparison between Picnic and Unlucky Man is complicated in terms of the number of false starts. The three measures of replacement, repetition and speech rate did not show any significant differences across the tasks. However, a trend is clearly seen with more repetition, reformulation and replacement being associated with the performances on the unstructured tasks.

Regarding accuracy, task structure has proved to have an effective influence on performance since performance on the structured tasks is significantly more accurate than performance on the unstructured tasks. The statistical analyses show that both structured tasks are significantly more accurate than the unstructured tasks. With regard to complexity, a significant difference is obtained between one of the structured tasks and the other three tasks, i.e. performance on Picnic is significantly more complex than performance on all other tasks. It is clear that, in general, the structured tasks have elicited more complex performances than the unstructured tasks. Interestingly, when the effects of task structure and language proficiency on complexity were compared, the results showed that the effect size of task structure on complexity, $\eta^2 = .153$, was greater than the effect size of language proficiency on complexity, $\eta^2 = .148$ (See Table 6.28).

From the general effect of task structure on performance, it can be concluded that structure has an immediate and noticeable effect on fluency and accuracy, whereas its effect on complexity is not so straightforward. This could be explained by referring to the notion of an existing competition for attentional resources between the goals of accuracy and complexity. In effect, this could support the argument put forward by Skehan (1998) and Skehan and Foster (1997) who propose that the two aspects of form, i.e. accuracy and complexity, compete with each other to consume more attentional resources. They report that for cognitively demanding tasks, participants can not pay equally sufficient attention to both accuracy and complexity. Further discussions of the effect of task structure on complexity measure will be presented in a later section in the current chapter.

The consistent effect of task structure on the language performance of 80 test-takers of English in the current study demonstrates a clear contrast with the results some researchers have reported in their studies before. In contrast with a number of studies carried out in pedagogic situations (Foster and Skehan, 1996; Robinson, 1995, 2001; Skehan and Foster, 1997, 2001; Wigglesworth, 1997, 2001), Iwashita et al (2001) and Elder et al. (2002) working in an assessment context, claim that task characteristics and conditions have no direct effect on the test-takers' language performance. They argue that the results obtained from their research are consistently different from those of the previous studies probably because of the differences between pedagogic and testing contexts. Nevertheless, the results of the present study, which has also been carried out in a testing context, provide a clear contrast with those of Iwashita et al. (2001) and Elder et al. (2002). The results of the present study clearly indicate that characteristics of a task, in the case of this study task structure, will influence different aspects of language performance. These results would also suggest that the effects of

task characteristics on language performance should be carefully studied before tasks are selected for pedagogic or assessment purposes.

7.2.2 Degree of Task Structure

An overriding focus of this study was to investigate the influence of degree of task structure on language performance. Within a framework proposed in SLA, task structure was systematically defined, on the basis of which four tasks with varying degrees of structure were selected. Structured tasks contained either a problem-solution structure or a schematic sequential organization with a clear time line underlying the events that occurred in each task. Unstructured tasks, on the other hand, did not have a macrostructure, a clear time line or a fixed sequence of events. It was hypothesized that the effects of task structure on language performance would be in line with the degree of structure that tasks present. In other words, performance on structured tasks would be progressively, as a function of degree of structure, more fluent and more accurate than performance on unstructured tasks. However, for the purpose of analysis and discussion, the concept of progression should be considered and interpreted at two levels: general and detailed.

First, progression should be considered in terms of whether the structure of a task, in the sense of a contrast between the two groups, would generate progressively more fluency and accuracy in performance on the structured tasks. Second, progression should be viewed as whether the degree of structure, as a distinction between the tasks within the structured and unstructured categories, has influenced performance. Therefore, progression in the first sense is general and would refer to more fluency and accuracy between the two categories of structured and unstructured tasks (i.e. have the structured tasks elicited more accurate and/or more fluent performances than

the unstructured tasks?). Whereas in the second sense, progression is detailed and would explore whether one task within the structured or unstructured categories has elicited more accurate and more fluent performances than the other one (i.e. is performance on Football more fluent and/or more accurate than performance on Picnic?). Optimally, a general progression would show that the Football and Picnic tasks have encouraged more fluency and/or more accuracy than the Unlucky Man and Walkman tasks have. In addition, a detailed progression indicates that a problem-solution structure has generated more fluency and accuracy than a schematic sequential structure, i.e. performances elicited by Football are more fluent and more accurate than those elicited by Picnic; and performances elicited by Unlucky Man are more fluent and more accurate than those elicited by Walkman.

The effects of degree of structure on fluency receive broad confirmation from the analysis of the data. Mean scores of total silence and number of pauses confirm that performance on structured tasks is progressively, both in a general and a detailed sense, more fluent than performance on unstructured tasks. Furthermore, a number of other measures of fluency confirm the existence of a general progression in fluency of the performances. In fact, fluency has constantly increased from the unstructured tasks towards the structured tasks. Figures 7.1 to 7.6 below show the mean scores of the fluency measures across the tasks. A detailed discussion of the repair fluency measures will be presented later in the current chapter.

Mean scores for the measure of accuracy also show a clear general progressive trend in the accuracy of the performances elicited by tasks as a function of the degree of task structure. That is, Football and Picnic have elicited more accurate performance than Unlucky Man and Walkman. However, regarding a detailed progression for the accuracy measure, the figures for the two structured tasks stay very close to one

another (Football: M = .42; Picnic: M = .43). Similarly, the two unstructured tasks have identical means for the accuracy measure (Unlucky Man: M = .30 Walkman: M = .30). These results, therefore, indicate that in terms of accuracy of performance the degree of structure does not have a detailed progressive effect on tasks.

Figure 7.1: Number of Pauses across Tasks

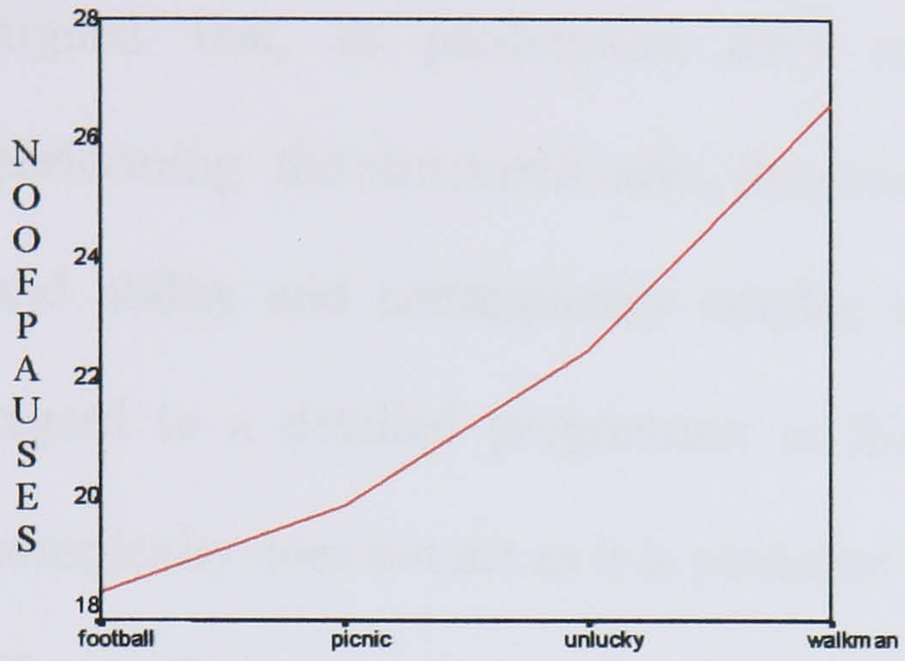


Figure 7.2: Total Silence across Tasks

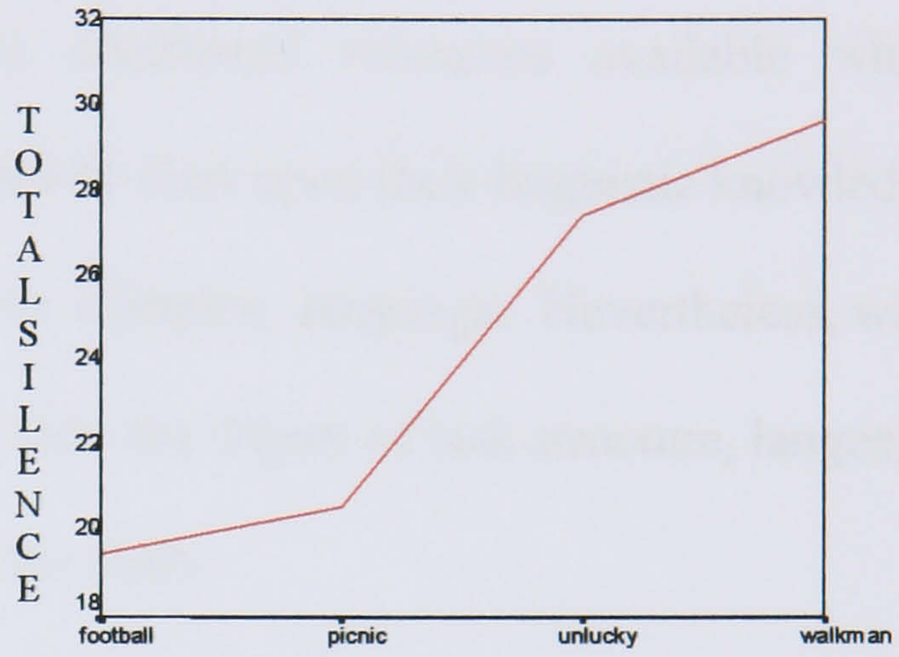


Figure 7.3: Pause Length across Tasks

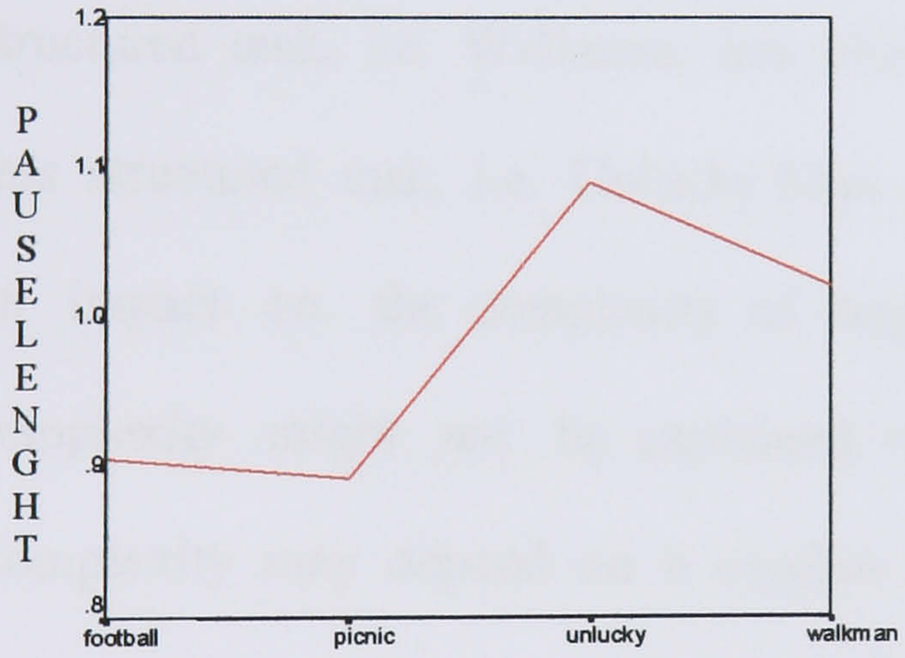


Figure 7.4: Length of Run across Tasks

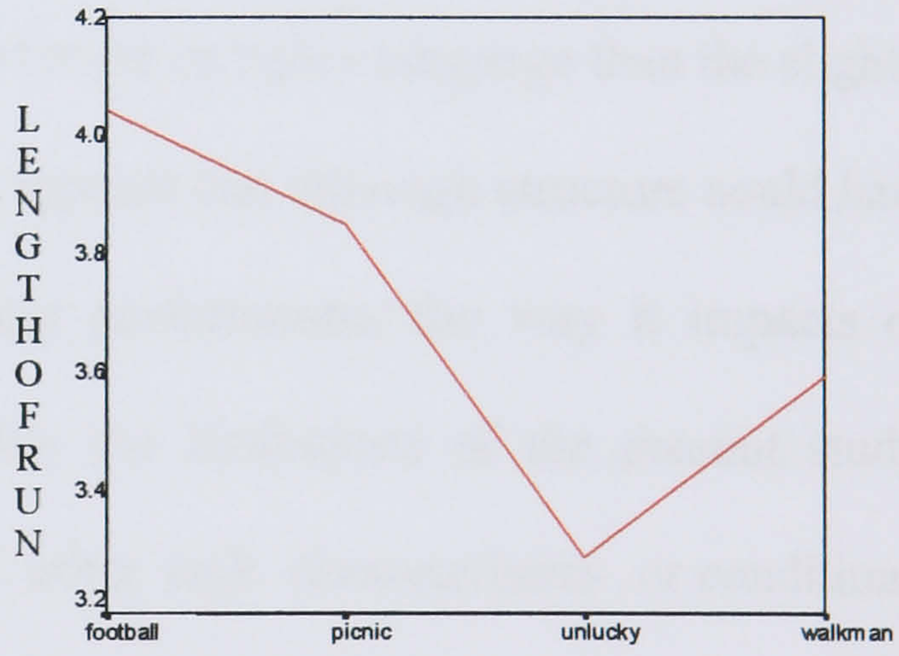
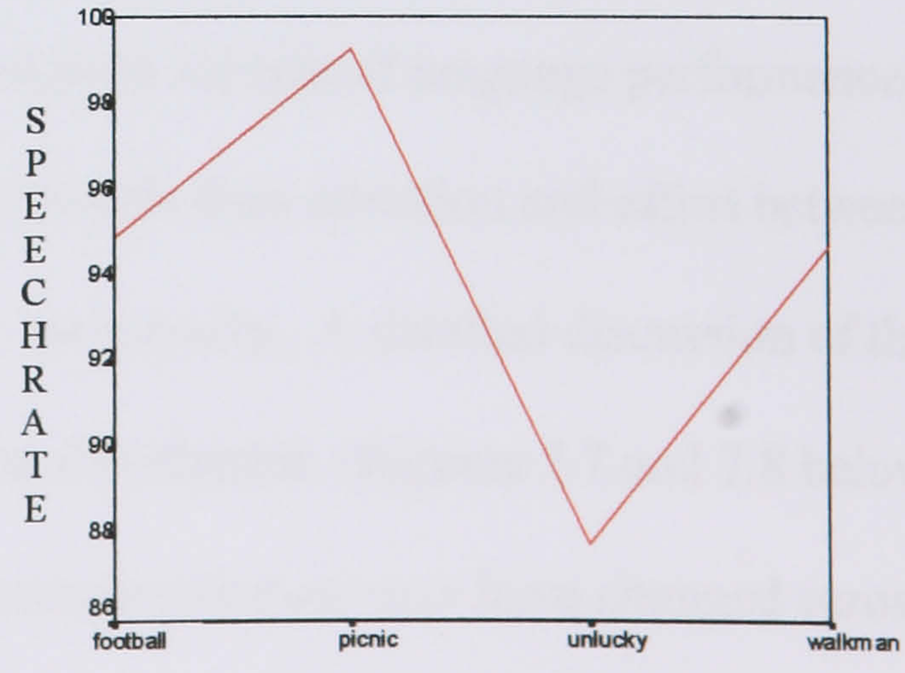


Figure 7.5: False Start across Tasks

Figure 7.6: Speech Rate across Tasks



The figures for complexity measures also provide a clear picture of a general progression from unstructured to structured tasks. The Picnic task presents the highest score (M = 1.59), i.e. it has elicited the most complex language, and the Unlucky Man has elicited the least complex language performance (M = 1.31). This clearly shows the contrast between the structured and unstructured tasks. It can be argued that, as participants have more attentional resources available while performing the structured tasks, they are able to draw upon their linguistic knowledge and ability and consequently employ more complex language. Nevertheless, with regard to a detailed progression in line with the degree of task structure, language complexity does not act as it is predicted in the study.

The performance elicited by the schematic sequential structure is significantly more complex than that elicited by the problem-solution structure. Similarly, the least structured task, i.e. Walkman, has elicited more complex language than the slightly less structured task, i.e. Unlucky Man. It appears that although structure could have an impact on the complexity of language performance, the way it impacts on complexity might not be explained within the limitations of the present study. Complexity may depend on a number of other task characteristics or conditions, which require further investigation. Another argument may concern the tradeoff between the two aspects of form, i.e. accuracy and complexity (See Chapter II section 2.5.3 for a detailed discussion of tradeoff between aspects of language performance). It can be argued that language learners would divide their attention and effort between the priorities they give to accuracy or to complexity. A detailed discussion of the complexity measure will be presented later in this chapter. Figures 7.7 and 7.8 below show how the mean scores of accuracy and complexity measures have changed across the four tasks. Figure 7.7 clearly demonstrates that there is a general progression, in

terms of the degree of structure, across the tasks for the measure of accuracy. It is worth noting that, as explained earlier in this chapter, a general progression refers to the progressive increase of the measure between the two groups of tasks, i.e. the structured versus the unstructured tasks. However, Figure 7.8 suggests that with the complexity measure a clear progression, in terms of the degree of structure, is not seen.

Figure 7.7: Accuracy across Tasks

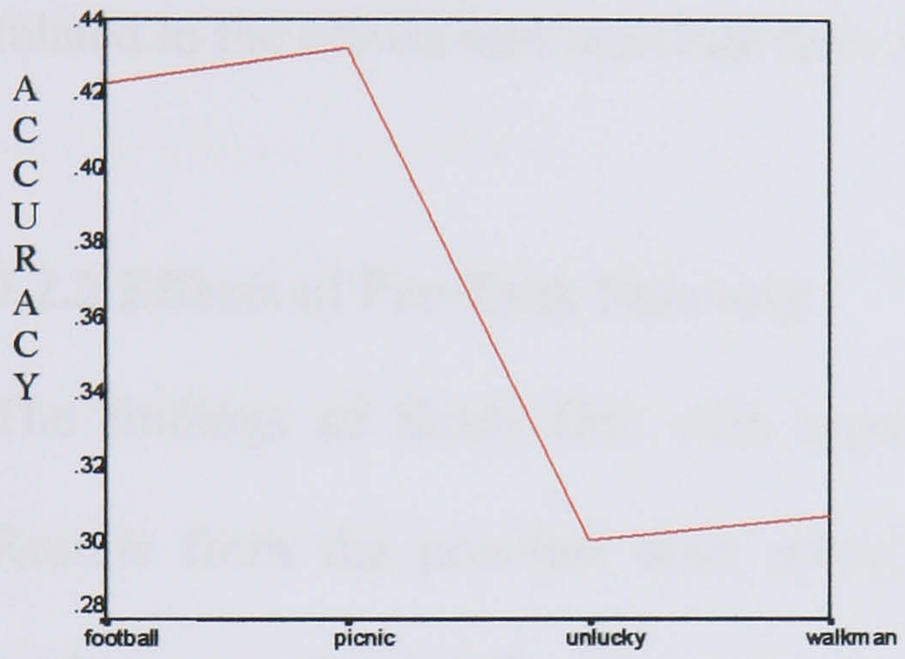
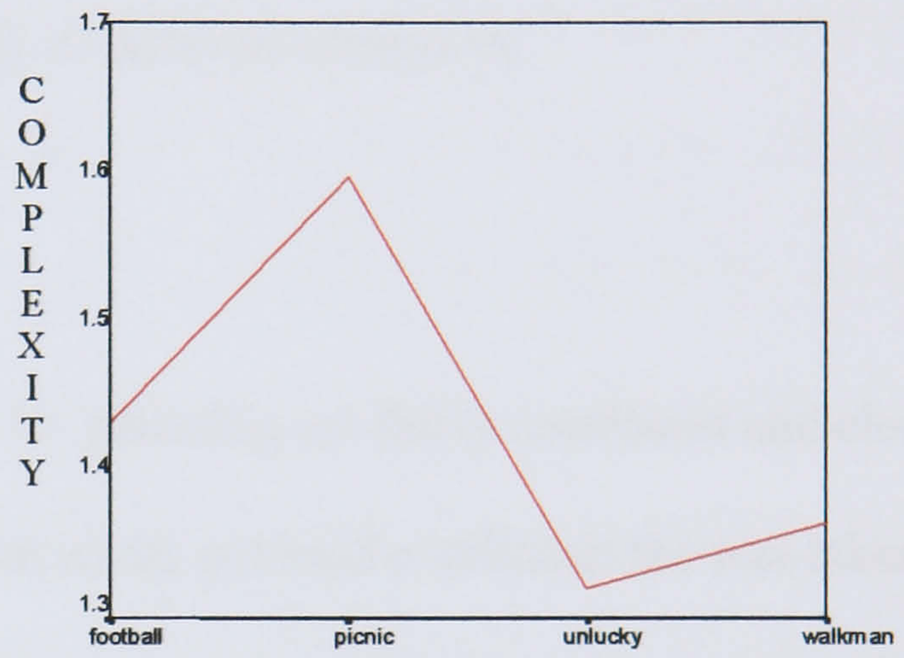


Figure 7.8: Complexity across Tasks



The last issue to be discussed regarding the effects of task structure is the significance of the effect size obtained in different statistical analyses in Study One. The measure of effect size is continuously being considered in SLA research since the reliability of significant findings is frequently questioned if adequate information about the effect size is not provided (Fulcher & Marquez Reiter, 2003). In fact, effect size describes the amount of the total variance in the dependent variable that is predictable from the knowledge of the levels of the independent variable. Values for eta squared can range from 0 to 1. Different researchers have provided different benchmarks to assess the largeness of the effect size. Pallant (2001) recognizes a value of .01 as a small, .06 as a moderate and .14 as a large effect size. Tabachnic and Fidell (1996) have considered an effect size of .04 as a small, .13 a modest and .71 a very large value.

The results of the ANOVAs on task structure (Table 6.15) show the significant differences for accuracy, complexity, length of run, total silence and proportion of time spoken across the tasks. More importantly, the results indicate that the eta squared for all these measures are noticeable (ranging from .151 to .267). As the effect size represents the proportion of variance of the dependent variable that is explained by the independent variable, the eta squared figures in this study suggest that the total variance in the above-mentioned measures is to a considerable extent related to the effects task structure have had on different measures.

7.2.3 Effects of Pre-Task Planning

The findings of Study One with regard to planning are fairly consistent and clear. Results from the post-hoc tests reveal that under planned conditions the test-takers' performance is significantly more fluent, more accurate and more complex than performance produced under unplanned conditions. These results broadly confirm the findings of Foster and Skehan (1996), Mehnert (1998), Ortega (1999), Skehan and Foster (1997) and Wigglesworth (2001). However, these results are in distinct contrast to Iwashita et al. (2001) and Elder et al. (2002) who reported that planning conditions have no effect on candidates' performance on tasks in an assessment setting. The results of the data analysis in the current study verify that pre-task planning provides the test-takers with an opportunity to focus on form as well as on meaning and enables them to draw upon their knowledge and skills to produce significantly more fluent, more accurate and more complex language.

Various measures of temporal fluency, i.e. total amount of silence, length of run, pause length, proportion of time spoken and speech rate have significantly improved under planned conditions. The number of pauses has also greatly reduced under the

planned conditions. However, the improvement in this measure does not reach statistical significance. In addition, measures of accuracy and complexity have been significantly higher under planned conditions, indicating that pre-task planning is effective in improving test-takers' accuracy and complexity. Although the results for complexity show significant differences resulted from the effect of pre-task planning, the effect size for complexity measure is a small one ($\eta^2 = .023$). This does suggest that although pre-task planning has influenced performances with greater complexity, the variance created in the complexity of the performance is not hugely influenced by pre-task planning. Finally, planning conditions have not strongly influenced different measures of repair fluency (to be discussed later). Surprisingly, each of the repair fluency measures appears to be higher under planned conditions. Figures 7.9 to 7.16 show the mean scores for the total amount of silence, number of pauses, length of run, mean length of pauses, proportion of time spoken, speech rate, complexity and accuracy under both planned and unplanned conditions for all the four tasks.

Figure 7.9: Total Silence under both Planning Conditions

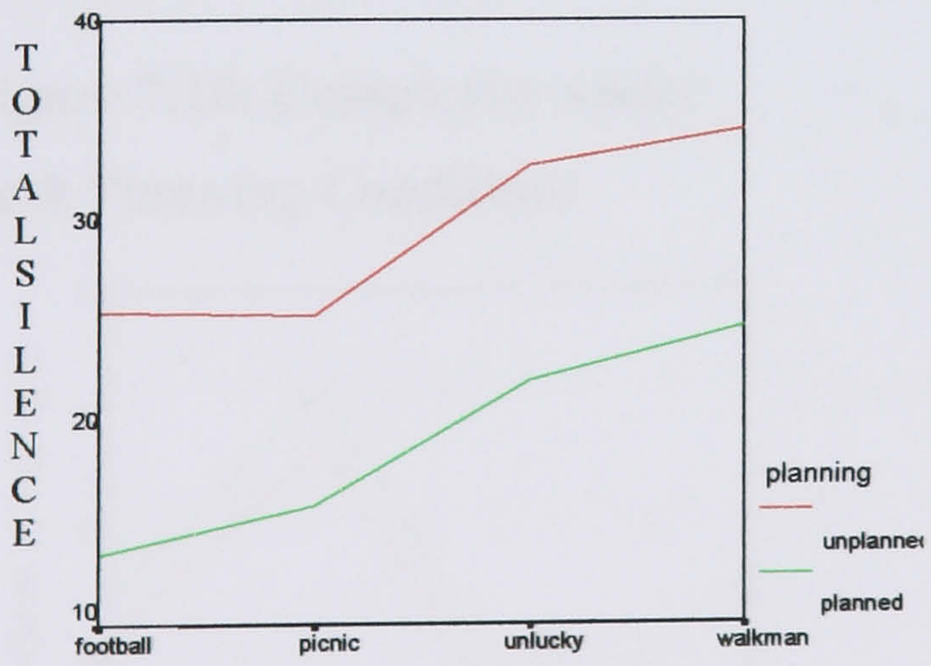


Figure 7.10: Number of Pauses under both Planning Conditions

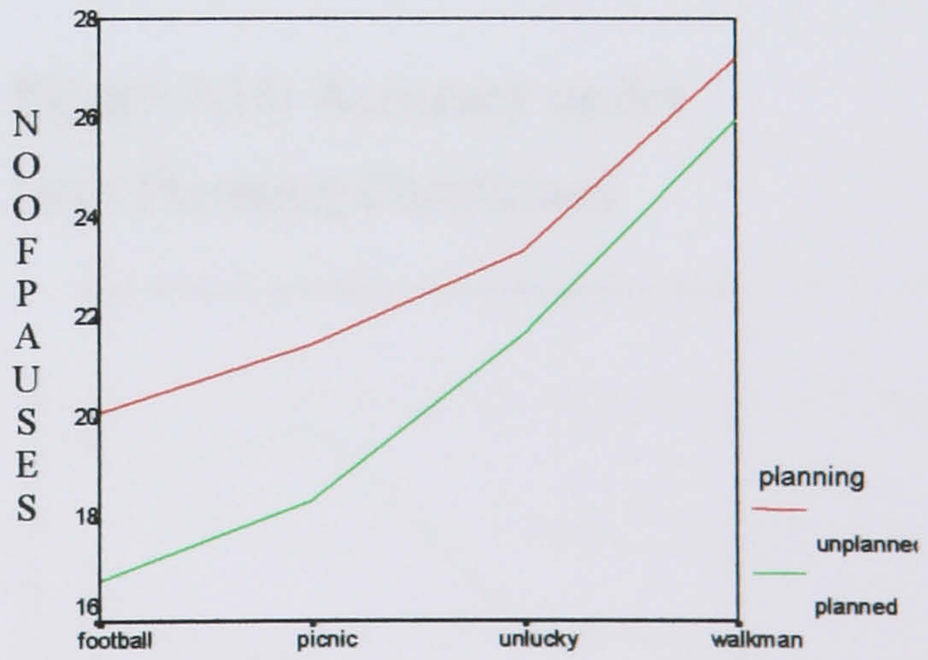


Figure 7.11: Length of Run under both Planning Conditions

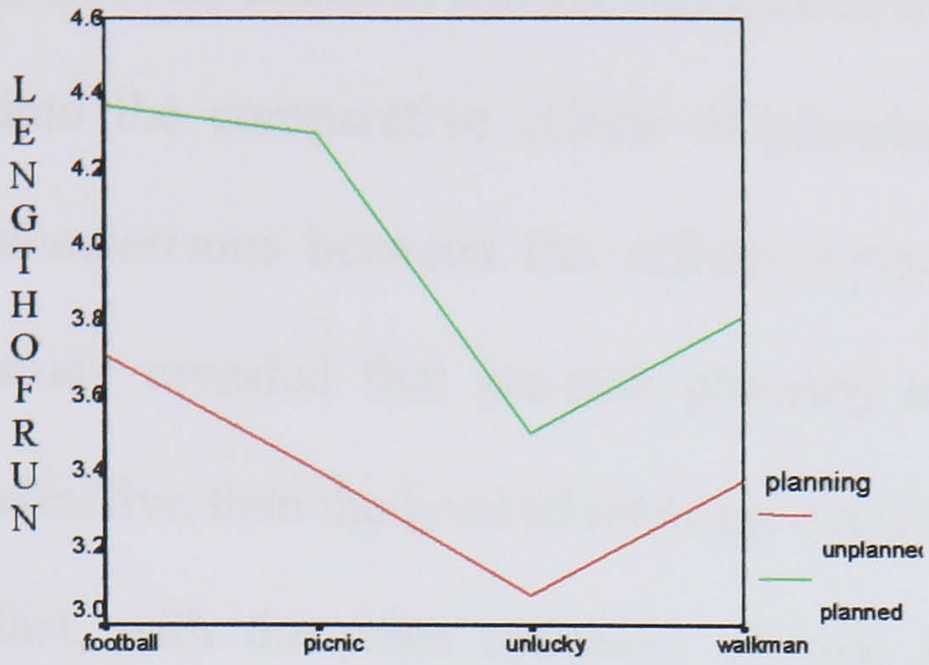


Figure 7.12: Pause Length under both Planning Conditions

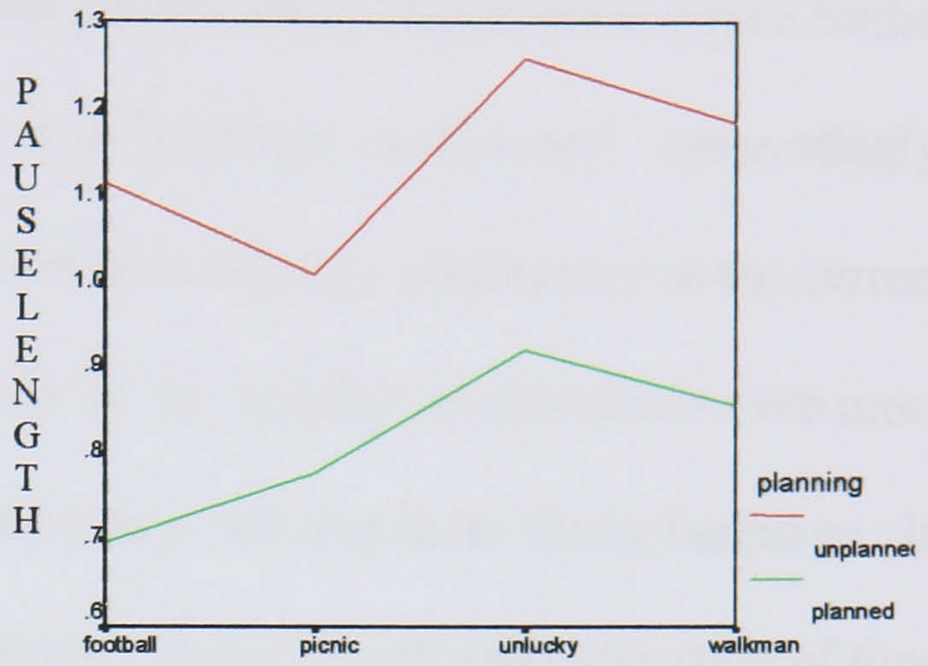


Figure 7.13: Prop. Time Spoken under both Planning Conditions

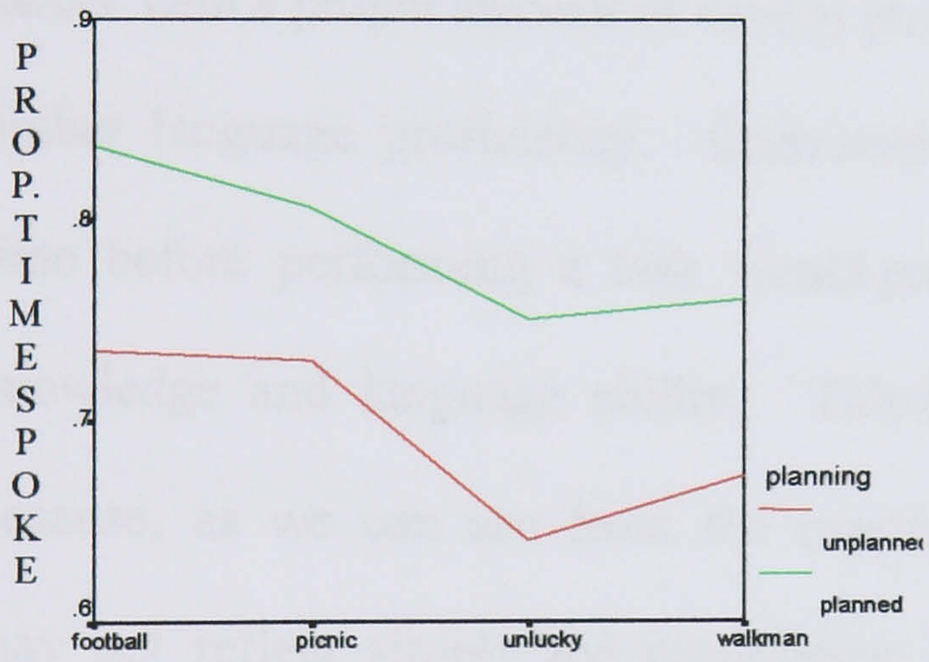


Figure 7.14: Speech Rate under both Planning Conditions

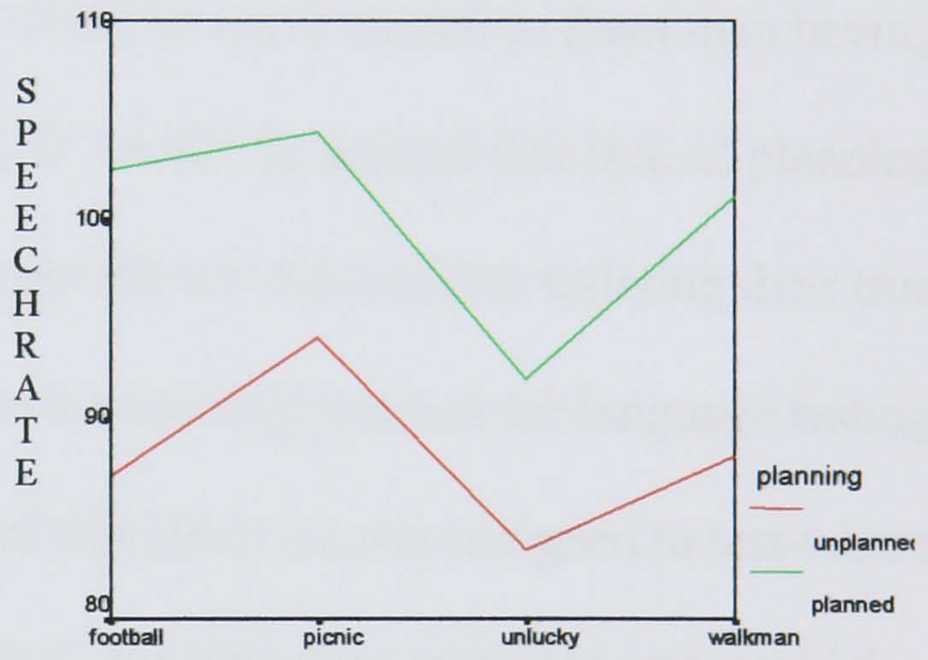


Figure 7.15: Complexity under both Planning Conditions

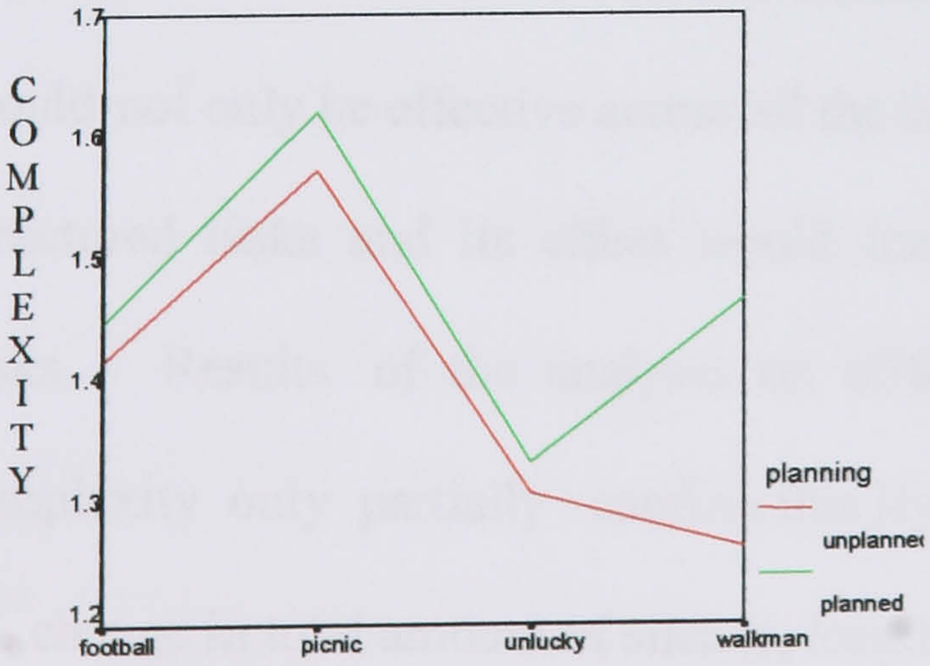
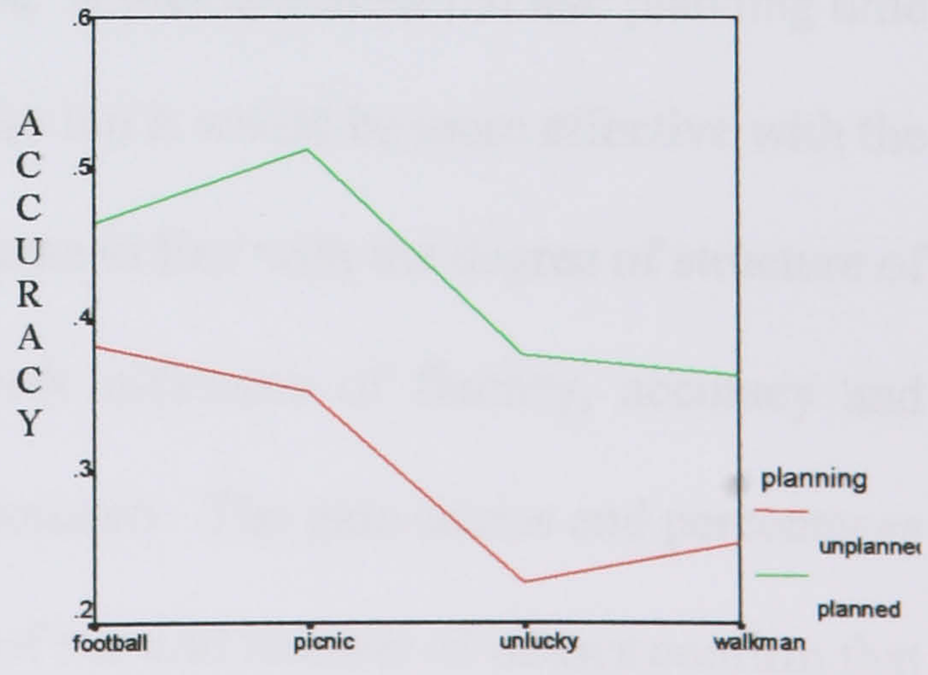


Figure 7.16: Accuracy under both Planning Conditions



As mentioned earlier, the effects of planning conditions on the performance of L2 learners have been repeatedly discussed in SLA research over the past two decades. However, different analyses conducted in the present study enable me to probe further into the comparative effects of planning and language proficiency. Interestingly, comparisons between the effects of planning and language proficiency in the current study revealed that pre-task planning tends to be equally or sometimes even more effective than the level of language proficiency in producing more fluent language. In fact, with the three measures of total silence, pause length and proportion of time spoken, the test-takers have benefited more from having planning time than from having a higher proficiency level. It could then be argued that providing the test-takers with a proper amount of time to plan could be more crucial to them than having higher language proficiency. Conversely, it could be argued that lack of planning time before performing a task would prevent the test-takers from utilizing their true knowledge and language ability. This has crucial implications for language testing because, as we can see from the results of this study, scores assigned to test-takers may not reflect simply the proficiency level, but also the conditions under which a task is performed.

Another salient role of planning in the present study was predicted with regard to the interaction it would have with task structure. It was hypothesized that planning time would not only be effective across all the tasks but it would be more effective with the structured tasks and its effect would increase in line with the degree of structure of tasks. Results of the analyses on different measures of fluency, accuracy and complexity only partially confirm this Hypothesis. The gain scores and percentages of change in total amount of silence, length of run and number of pauses confirm that planned participants were more fluent on the structured tasks. For mean length of

pauses and speech rate, the Football task has elicited the most fluent language under the planned condition. However, the results of other measures of fluency, accuracy and complexity are mixed and only partially confirm the Hypothesis.

Regarding accuracy, the Unlucky Man task shows the greatest improvement (72%) and the Football task demonstrates the smallest improvement (23%) under planned conditions. Regarding complexity, the Walkman task has the greatest percentage of change (15%) under planned conditions, whereas the Football and Unlucky Man tasks have had the smallest change (2%). The results suggest that although all tasks have been influenced by the planning conditions, there is no clear progression, as a function of task structure, in accuracy and complexity of performance. In fact, it appears that participants have employed planning time to increase the fluency of their performance in general. But with accuracy and complexity there is not enough evidence to speak about a clear pattern of attention allocation. It could be argued that tasks with varying degrees of structure are not, to the same extent, affected by planning conditions. In effect, unstructured tasks appear to be more demanding to the test-takers and this might affect the amount of effort they put into performing the task more effectively. Another argument is that there might be other characteristics in the selected tasks which are influencing accuracy and complexity of performance but are not accounted for in this study.

In earlier chapters it was argued that the structure of a task greatly reduces the cognitive load that is imposed on learners or test-takers. This reduction in the cognitive demands of a task, in turn, provides the test-takers with a better opportunity to focus on both form and meaning simultaneously. The results clearly showed that when pre-task planning is available, the test-takers are more likely to employ their linguistic resources more effectively in both the structured and the unstructured tasks

and produce more fluent and more accurate performance. Although complexity of performances is increased under planned conditions, the effect size measure indicates that planning has a small influence on the complexity of performances. This suggests that there might be other task characteristics and/or conditions that have influenced performance on tasks in a way that more complex language is elicited.

7.2.4 Effects of Language Proficiency

The findings of this study for language proficiency are straightforward, and provide confirmation that the data elicitation format has produced results consistent with both the institutional test and Oxford placement test. The performance of the participants at the intermediate level proved to be consistently more accurate, fluent and complex than those at the elementary level. It was hypothesized that high proficiency test-takers would benefit more from planning time, in terms of fluency, accuracy and complexity than would the low proficiency test-takers. The gain scores and the percentage of change in the fluency measures of the high and low-proficiency groups show that the low-proficiency test-takers have used pre-task planning more effectively in many of these measures. For total amount of silence, length of run, number of pauses and speech rate the larger percentages of change belong to the low-proficiency group. This suggests that fluency could be strongly influenced by performance conditions rather than by the effect of language proficiency. In contrast, regarding the accuracy and complexity measures, high-proficiency test-takers have benefited more from pre-task planning. This indicates that, having a potentially higher language ability, high-proficiency test-takers have more resources to draw upon and employ the planning opportunity more effectively in producing more accurate and more complex language. In addition, the effect size for accuracy and complexity measures revealed

that the significance reached for these measures are to a great extent influenced by task, planning and proficiency levels.

Taking the effect of language proficiency into consideration, it was also hypothesized that under the planned condition, high proficiency test-takers would perform better than low-proficiency test-takers on unstructured tasks. However, the results of different measures of fluency, accuracy and complexity are very mixed, indicating no particular trend relevant to the degree of structure for high-proficiency test-takers under planned conditions. In effect, for total amount of silence the two structured tasks are performed with greater improvement in their percentage of change, suggesting that the high-proficiency group take advantage of both planning time and their language ability to do the less cognitively demanding tasks better. In contrast, for false starts, performance on the unstructured task tends to be better than that on structured tasks. Regarding the accuracy and complexity measures, one of the unstructured tasks receives a surprisingly high percentage of change, whereas one of the structured tasks shows the smallest percentage of change.

These results are generally different from those presented by Wigglesworth (2001) who reported that high-proficiency candidates performed better on the unstructured tasks when they were given the opportunity to plan. However, the present study is different from that of Wigglesworth (2001) regarding both the underlying assumptions and the operationalization of structure and planning time. It can be argued that the variations between the results of the two studies may be connected to the fact that structure is investigated in a more systematic way in the present study. Moreover, it should be noted that there is a noticeable difference between the two studies in terms of the proficiency levels of the participants.

7.2.5 Perceptions of Task Difficulty

One main objective of the study was to explore the test-taker perceptions of task difficulty as a function of task structure. It was hypothesized that perceptions of task difficulty would be in line with the predicted difficulty of tasks, in terms of task structure. Results of the analysis of the responses to the retrospective questionnaires revealed that the test-takers, as predicted, rated the unstructured tasks as significantly more difficult than the structured tasks. In effect, the Football and Picnic tasks are rated as less difficult than the Unlucky Man and Walkman tasks, with little difference between the pair of tasks in each case. More importantly, the results of three-way ANOVA showed that the effect size of the task structure on ratings of perceptions is noticeable, suggesting that this significant difference is reached as it is remarkably influenced by task structure. Non-planners rated tasks generally as more difficult than did the planners. Furthermore, participants rated unstructured tasks as more difficult under the unplanned conditions than under the planned conditions. However, no effect was found that could be attributed to their language proficiency or interaction among the dependent variables. It is particularly interesting that the lowest difficulty rating was given to the Picnic task under the planned conditions. This will necessarily require further research, because Picnic under the planned conditions has elicited performances with high levels of accuracy, complexity and fluency.

The results of the questionnaires on the perceptions of task difficulty generally confirm the findings of Robinson (2001) who found task complexity would affect learner perception of task difficulty. Surprisingly, these results do not agree with Elder et al. (2002) who report that test-taker perceptions of difficulty were not related to the difficulty of performance conditions of the tasks predicted in their study.

A second section of the planned questionnaires explored the participant perceptions of the usefulness of planning time for different tasks. However, no significant difference was observed among the ratings of usefulness they have assigned to different tasks. The lack of a significant result in this case might raise a number of different issues about the participants and the conditions. It could indicate that although the participants' perceptions are affected by task difficulty, they could not evaluate themselves in terms of the help they have received from the planning time. Another argument is that, as planning was a between-participants variable, only half of the participants had the opportunity to plan and thus had to answer the questions in relation to the usefulness of planning. As a result, the test-takers who have answered this section of the questionnaire might have not had any idea of comparing their situation with someone who has performed the same tasks under unplanned conditions.

The open-ended part of the questionnaires required the participants to add any comments they had about the four tasks or about the test. Only a few of them did actually respond to this question. As there was not adequate evidence, finding a clear conclusion seems to be difficult.

The general results of the analysis of participant perceptions confirms that task difficulty, i.e. lack of structure, not only affects the performance of the test-takers but also directly influences their perceptions of task difficulty. In effect, it can be concluded that the participants have a clear insight into whether particular task characteristics and conditions would make a task easier or more difficult to perform. More importantly, these findings have important implications for task-based syllabus design, task-based instruction and task-based language assessment. Although syllabus designers and test developers cannot merely rely on student ratings of difficulty as the

criterion for grading, selecting and sequencing tasks, the findings of this study indicate that they could receive great assistance and feedback from the test-takers' perceptions of task difficulty.

7.3 Conclusions

The present study set out to determine whether task characteristics and performance conditions have any effect on the language performance of 80 Iranian L2 learners of English in an assessment setting. The salient purpose of the study was to define task structure in a systematic way and consequently to explore if degree of task structure would affect performance on oral narrative tasks. The need expressed by SLA researchers for investigating different characteristics of tasks in an assessment context (Chalhoub-Deville, 2001; Norris et al., 1998; Skehan, 1998, 2001; Wigglesworth, 2001) and the inconsistent results reported by some researchers in the task-based assessment context (Elder et al., 2002; Iwashita et al., 2001) were the prime motivations of designing Study One. In effect, the current study attempted to investigate the effect of task structure, pre-task planning condition and proficiency level on test-takers' language performance, and to explore their perceptions of task difficulty in a testing context.

The findings of the study, resulting from a systematic investigation of task structure and planning time, have provided more support for the recent progress in cognitive theories of SLA. It is now clear that availability of task structure and pre-task planning has a facilitative impact on the learners' performance. Inherent structure of a task, whether in the form of problem-solution or schematic sequential organization, reduces the cognitive load of a given task and thus frees up attentional resources for the participants to attend to different aspects of form and meaning. In addition, under

planned conditions the communicative stress of the task is reduced and thus there is more space for the test-takers to assess task demands and to employ linguistic or strategic resources available to them to perform the task better.

In summary, these findings could broaden perspectives for SLA researchers and pedagogues as well as language test-developers to better realize what effects task characteristics might have on language performance in task-based instruction and task-based assessment. The results of the current study evidently show that L2 performance greatly varies when different characteristics and conditions of tasks are being manipulated. These findings also suggest that task difficulty, despite its intricate nature, can be explored to be adapted for various teaching and testing purposes. Hence, the results of this study should help test-designers in their selection and operationalization of tasks in order to develop tests which are of appropriate difficulty to the candidates and could be "less impositional and more humanistic" (McNamara, 2000). These results could further provide syllabus designers with more insight into the selection, grading and sequencing of the tasks in task-based syllabi. The findings of the present study could eventually contribute to the understanding of L2 teachers when they employ different tasks for both their teaching and classroom testing purposes.

7.4 Implications for Further Research

The results of the current study have demonstrated that task structure, as hypothesized, has had discernible effects on the language performance of Iranian test-takers of English. With regard to fluency and accuracy, the results are very clear and support the hypotheses of the study suggesting that task structure would reduce the cognitive load of the task and provide the learners with an opportunity to focus on

fluency and accuracy. Regarding syntactic complexity of performance, however, results suggest that task structure does not have a clear impact on performance. It could be argued that other task characteristics appear to be influencing the syntactic complexity of L2 performance. Therefore, investigating what characteristics of oral narrative tasks would influence the syntactic complexity of performance is a clear line of inquiry to be pursued in the current research.

As results on fluency measures suggest, no regular pattern is observed with regard to the repair fluency measures. In effect, although task structure and pre-task planning have had a considerable influence on temporal measures of fluency, a clear impact is not observed for reformulations, repetitions and replacements. Hence, more investigations are required to explore different aspects of repair fluency measures and to investigate what characteristics and conditions would impact on them. Furthermore, the results of the break down fluency measures, i.e. number of pauses and length of pauses, showed that there were many pauses of longer than .4 of a second in the performance of second language learners. However, it is not clear whether these long pauses occur at clause boundaries or whether they mainly interrupt performance in the middle of clauses. Hence, more investigations are needed to find out where these interruptions occur. These are some of the significant issues the results of Study One have raised. Undoubtedly, more systematic research is required to answer such questions.

CHAPTER VIII

Research Design: Study Two

8.1 Overview

Study One, as explained before, essentially set out to determine whether different task characteristics and task conditions would influence language performance on tasks. The results of the statistical analyses of the data clearly demonstrated that pre-task planning has a considerable overall effect on certain aspects of fluency, accuracy and complexity of the language performance of Iranian language learners. Furthermore, the results of Study One indicated that task structure influenced accuracy and fluency of language performance on oral narrative tasks. However, the results did not show any direct effect of task structure on complexity of performance. Based on these findings, Study Two is developed primarily to investigate:

- What task characteristics would influence the structural complexity of the language performance?
- What interaction would there be among different task characteristics which could influence performance?

As mentioned in Chapter V complexity of performance in Study One was measured in terms of the ratio of subordination of each performance. Following Foster et al. (2000), all the transcribed data were marked for AS-units and subordinations. Then, the ratio of the subordinate clauses to the total number of clauses in each performance was calculated. This ratio was the measure that represented the structural complexity

of the performance of a test-taker on each of the tasks. The means of complexity across the four tasks in Study One showed that Picnic elicited the most and Unlucky Man the least complex performance (means of the structural complexity for Walkman: 1.36, Unlucky Man: 1.31, Picnic: 1.59, Football, 1.43). With the structured tasks, Picnic elicited more complexity than Football, and with the unstructured tasks, Walkman elicited more complexity than Unlucky Man. However, the structured tasks, i.e. Football and Picnic, generally elicited more complexity than the unstructured tasks, i.e. Walkman and Unlucky Man. In Study One, it was hypothesized that, as the presence of structure in a task would ease the cognitive processing load of a task, the structured tasks would elicit more complex language. But the results indicate that in the structured tasks Football elicited less complex language than Picnic. In the unstructured tasks, Walkman elicited more complex language than the Unlucky Man. Therefore, a main implication arising from the findings of Study One is that the impact of task structure on complexity of performance has yet to be clearly identified. This motivated a further enquiry into exploring other task characteristics that influence performance in general and complexity of performance in particular. Hence, the main research question in Study Two will be what task characteristics would influence language performance in tasks in such a way that more syntactic complexity is elicited.

In order to form a hypothesis based on this question, a detailed investigation of the data was initially made. All the transcripts of the data from Study One were investigated. The careful inspection of the data revealed that in the performances on the Picnic and Walkman tasks more subordinating clauses were elicited when the participants attempted to relate two co-occurring events or tried to conjoin a main event with other events happening in the background of the picture stories. In so

doing, in fact, it seems that the participants were trying to include the details of the events that occurred in the background in order to support, elaborate or assist narrating the main events of the story. Examples from the data transcripts of performance on Picnic and Walkman containing subordinating clauses are provided here. The subordinating clauses are italicized in these examples (See Appendix 4 for the coding symbols and Appendix 5 for samples of the coded data).

Picnic:

Transcript 1: | *when er they open the basket :: they saw the dog :: jumping out of the coming out of the basket* |

Transcript 2: |*and at the time they were do this :: their mother called he them ::and told them :: that where they should go* |

Transcript 3: | *and when they go to em eat their lunch :: their dog they figure out :: that their dog has eaten all the food* |

Walkman:

Transcript 1: |*and he do not realize and do not understand :: what happened around them around him* |

Transcript 2: | *and er in picture four er he was crossing again he was crossing another street :: that a policeman er arres- arrested er arrested other em thiefs* |

Transcript 3: | *when was crossing the street :: a car was a car was gonna to have an accident with with him* |

While investigating the transcripts of the data, the sets of pictures in each of the picture stories were further carefully evaluated and the relationship between the syntactic complexity of performance in the data and the events occurring in the picture stories was carefully reconsidered. In two of the tasks, Walkman and Picnic, it is clearly seen that many actions and events happen in the background, which relate to

or combine with the main events of the story. In Picnic, a number of events happen in the background, which can be meaningfully incorporated into the main story. These background events seem to change the ongoing story as well as the outcome of the story. In Walkman, many events are happening in the background along with the events occurring in the main story. However, as Walkman does not have a clear inherent structure, the background events do not appear to be meaningfully incorporated in or relate to the main events of the story. The other two tasks, Football and Unlucky Man, are stories explicitly based on the main events. In Football few, if any, events occur in the background. In Unlucky Man, the main events form the story and nothing happens in the background.

Investigation of the data transcript and inspections of the picture stories has greatly contributed to formulating a hypothesis for Study Two. However, before forming the hypothesis, a review of SLA literature is required to provide an appropriate theoretical framework for the hypothesis. Moreover, a review of the wider SLA literature would supply this research with a deeper insight into the relevant characteristics of a task, which may have not been discussed in task-based research literature.

First, in order to have a broader perspective towards the different characteristics of picture stories, the impact of visual stimuli on language performance has to be considered here. In the second language assessment literature, it is established that visual stimuli would influence learners' comprehension of verbal communication while they are involved in L2 conversations and interactions (Ginther, 2002). A substantial body of literature also discusses the effects visuals would have on written text comprehension (Mandle & Levin, 1989; Winn, 1991). However, this literature has not led to a consensus or an overarching theoretical framework with respect to the interaction of visual and textual sources of information (Ginther, 2002). As regards

the current study, however, it is hoped that the results would reveal more about the way the visual stimuli interact with L2 learners' comprehension and communication. Second, in order to know more about different characteristics of narrative picture stories, the relevant literature on narrative discourses and performance on narrative picture stories are to be studied. This literature shows that the discourse of second language learners while they work on narrative picture stories has been carefully investigated by a number of researchers (Dry, 1983; Reinhart, 1984; von Stutterheim, 1991). Some studies have attempted to find the effects of the organizational structure of narratives on second language learners' spoken discourse. In these studies, the idea of foreground and background information appears to form a significant part of how the organization of narratives is defined. Polanyi-Bowditch (1976) initially introduced a basic definition of foreground and background in a narrative:

> Narrative is composed of two different kinds of structures: temporal structure, which charts the progress of the narrative through time by presenting a series of events which are understood to occur sequentially; and durative/descriptive structure, which provides a spatial characterological and durational context for which the temporal structure marks time and changes of state. (p. 61)

Following Polanyi-Bowditch, Hooper and Thompson (1980) attempted to provide a more comprehensive definition of foreground and background as an overall characteristic of spoken discourse. They argued that foreground and background were not mainly restricted to narrative discourse, but are a characteristic of other genres as well. They contended that, in any speaking situation, some parts of what is said are more relevant or enjoy more significance than others. *Background,* as defined by Hooper and Thompson (1980), is the part of discourse which merely assists, amplifies, or comments on it, whereas, *foreground* is the material which supplies the

main points of discourse. Tomlin (1984) considers the concept of foreground and background information with regard to the propositional values they bear on texts. As he argues, foreground information is used in the analysis of text artifacts to describe those propositions in the text which are more important or central to the development of the overall discourse theme. He adds that background information is used to describe those propositions which elaborate or explicate the foreground information. Turner (1992) has studied the effects of the narrative picture stories on enhancement of the learners' thinking patterns and their understanding of the effective use of language. Focusing on the visual aspects of narrative picture stories, she argues that, apart from the time sequence that carries the main theme of a narrative forward, there are other elements such as cause and effect or descriptions that help the complementary actions in a narrative develop. Turner (1992), in effect, is taking the time sequences as the main theme, i.e. foreground, and the other elements as complementary, i.e. background. She insists that presence of these 'other elements' in a picture story would provide learners with frameworks for organizing, sorting and storing information effectively and for making connections across contexts and texts.

Recently, Bardovi-Harlig (1998), has reconsidered the notions of foreground and background and defines them as two important parts of a narrative discourse. She explains that the foreground in a narrative relates events belonging to the skeletal structure of the discourse, which consists of clauses that move time forward. She states that "the temporal point of reference in any one event in the foreground is understood as following that of the event preceding it" (p. 475). Bardovi-Harlig (1998) argues that the background elements would not be narrating the main events of the narrative but would provide supportive materials that elaborate on and evaluate the events in the foreground. She contends that:

> For example, a background may contribute to the interpretation of an event by revealing a prior event (located before the narrated event on the time line), making a prediction about the outcome of an event (located after the event on the time line), or evaluating an action reported in the foreground (not located on the time line). (Bardovi-Harlig, 1998, p. 476)

In the present research, Bardovi-Harlig's (1998) notions of foreground and background will be adopted in formulating the hypotheses. By definition, therefore, foreground in a narrative story forms the story line and consists of the main events that move the time and events forward. On the other hand, background does not narrate the main events but provides support to the main events and consists of the scene-setting and scene-complicating elements. The term *"grounding"*, which is defined in the literature (e.g. Polanyi-Bowditch, 1976; Hooper and Thompson, 1980) as the linguistic features associated with the distinction between foreground and background, is employed in the current study to include the concepts of both foreground and background. The definition of grounding in this study is also expanded to include all the relevant information in a narrative picture story that refers to the foreground and/or background information provided. Thus, the term *"foreground task"* will be used to refer to picture stories which are reliant on foreground information. Picture stories that have both foreground and background information will be called *'foreground and background tasks'*.

To summarize, there are two task characteristics that Study Two will focus on. First, the effect of grounding on language performance in general and on syntactic complexity of performance in particular will be investigated. Second, the effects of task structure will be explored to find out whether Study Two replicates the findings of Study One and whether there are interactions between task structure and grounding.

Thus, based on the findings of Study One and drawing on the relevant aspects of SLA literature reported in this section, the following hypotheses are formulated.

8.2 Hypotheses

Hypothesis 1: Language performance in foreground and background tasks is more complex than language performance in foreground tasks. This hypothesis is based on the results of Study One in which language performance elicited by some tasks was more complex than performance elicited by other tasks.

Hypothesis 2: Language performance in structured tasks is more accurate than performance in unstructured tasks. This follows from Study One which draws upon Skehan and Foster (1999) and Wigglesworth (2001), who found that the performance in structured tasks is more accurate than the performance on unstructured tasks.

Hypothesis 3: Language performance in structured tasks is more fluent than performance in unstructured tasks. This follows from Skehan and Foster (1999) and Wigglesworth (2001), who found that the language performance in structured tasks is more fluent than the language performance in unstructured tasks.

It should be mentioned that the reason for including task structure, i.e. Hypotheses 2 and 3, in this study is twofold. Firstly, task structure is employed to enable the researcher to compare the findings of the current study with those of Study One. Second, task structure is included in Study Two so that it is possible to investigate whether task structure and grounding interact with each other and whether the interaction between these two characteristics influences learners' performance.

8.3 Methodology

8.3.1 Design

In this second study, the focus is on two characteristics of oral narrative tasks: grounding and task structure. This study attempts to find the effects of these two independent variables, i.e. grounding and task structure, on accuracy, fluency and complexity of language performance that is elicited by some oral narrative tasks. Therefore, a 2 x 3 factorial design is used with grounding and task structure as the two independent variables of the study. Grounding is the between-participant variable with two levels, indicating that half of the participants will perform tasks containing only background information and the other half will perform tasks which contain both foreground and background information. One purpose of the current study is, in effect, to compare participants' performance as influenced by the two levels of grounding. Task structure is the within-participant variable with three levels, i.e. all the participants will perform three tasks of different degrees of structure. The second purpose of the current study, therefore, is to compare the performance of the same participant on three different tasks with different degrees of task structure. The details of the design, tasks and participants will be provided in the sections that follow.

8.3.2 Tasks

In line with the theoretical assumptions of Study One and following the body of task-based language assessment research from which the motivations of the present research are drawn, oral narrative tasks are employed in the current study. As explained in Chapter III, narrative tasks are popular among different international testing organizations and are typically used as stimuli to elicit language performance (e.g. TOEFL's Tests of Spoken English). Narrative tasks in this sense refer to short

picture stories represented by a sequenced set of visual prompts which are shown to the participants while they are asked to narrate the story. Although the rationale for using oral narrative tasks in terms of their validity, reliability and authenticity of the tests is well established (For a detailed discussion see Chapter III), the prime reason for employing narrative tasks in the current research is to conform with the LT literature from which the theoretical assumptions of this research are drawn. In effect, I will use this type of task, i.e. oral narrative task, in my research because this type is typically employed by international testing organizations to assess test-takers' oral language ability.

Following the procedures taken in Study One, and in order to find suitable picture stories, both EFL and non-EFL sources were searched. The procedures, requirements, selection criteria and the type of the resources that were searched for narrative tasks have been explained in Chapter V, Section 5.3.2. The design of Study Two imposes two prime requirements on the selection of the tasks. First, three picture stories, each with a different degree of structure, are needed. In fact, in order to compare the results of the two studies regarding task structure, picture stories are required that can be placed in three categories: unstructured, schematic sequential and problem-solution. Definitions of all different types of task structure are given in Chapter IV. Second, picture stories that would show either only foreground information or both foreground and background information are required to be put in each of the grounding categories. In the event, the process of finding suitable picture stories for this study became more complicated, as the two task characteristics, i.e. grounding and task structure, had to be taken into consideration carefully. In effect, in the search of suitable picture stories, it was necessary to find a number of different tasks, each of

which demonstrating one certain type of structure with either of the grounding type of information.

The search for narrative picture stories in non-EFL materials was not successful since most of these materials either are too long or contain some linguistic cues, which would make them inappropriate for the purpose of the study. In selecting the picture stories, similar to Study One, a number of criteria were taken into consideration. The picture stories had to be of a suitable length, i.e. between 6 and 9 picture prompts, bear a clear story line, be interesting, culturally familiar and acceptable to the participants. The picture stories should be neither linguistically cued nor linguistically demanding. A total of 25 picture stories, which seemed to meet all the above-mentioned criteria, were initially collected from EFL teaching materials. Through further investigation of the picture stories with two experienced researchers[1], a total of 9 picture stories seemed to be appropriate in terms of the clarity and story line, the type of task structure they presented and the grounding criterion. The appropriateness of the picture stories regarding the cultural values was also considered. The details of the picture stories and their task characteristics will be presented under the relevant sections.

8.3.3 Task structure

As one purpose of the study is to investigate the effects of task structure on language performance and the interaction between task structure and grounding in a task, structured and unstructured tasks are needed. Following Study One, task structure is defined in terms of either problem-solution or schematic sequential organization

---

1 Once more, I would like to express my deep gratitude to Peter Skehan and Constant Leung for the invaluable time and professional comments they provided me with in selecting suitable picture stories.

structure. Definitions of both types of structure and the relevant discussions are presented in Chapter IV. Therefore, four structured tasks, two from each structure category, are needed to indicate problem-solution and schematic structure. However, as grounding is the second independent variable of the study, the picture stories would have to differ in the type of grounding they present. Similarly, two unstructured tasks, one from each grounding category, i.e. foreground versus foreground and background, are required. Following Study One, a picture story which lacks a clear structure in terms of its time line or macrostructure is considered as unstructured (See Chapters IV and V for a detailed discussion). In effect, an unstructured task refers to a picture story based on an arbitrary sequence of events in which many of the pictures can be ordered on any other sequence without the main theme of the story being changed. Table 8.1 shows the six tasks required for the design of Study Two.

Table 8.1

Task Characteristics and Tasks in Study Two

	Problem-solution	*Schematic Sequential*	*Unstructured*
Foreground	**Task 1**	**Task 2**	**Task 3**
ForegroundBackground	**Task 4**	**Task 5**	**Task 6**

There is a delicate point in the relationship between the two characteristics of task structure and grounding in a picture story. Task structure, as explained before, mainly refers to the timeline or the chronological sequence of the events. On the other hand, foreground information refers to the events happening in the story which move time forward. As the two characteristics deal with aspects of time in the story, there might be some overlap or interaction between the two aspects. However, the picture stories that are utilized in the current study are selected in a way that such overlap or interaction between the two characteristics is very marginal.

After receiving feedback from the results of the pilot study, four of the picture stories were eventually selected for the structured tasks category. Two of the picture stories, Keys (O'Conner, 1989) and Football, have problem-solution structures. In both stories, a problem occurs which affects the main characters of the story. The characters decide to use their initiatives to solve this problem. While they are involved in solving the first problem, other problems happen and they consequently think of further solutions. Finally, there is a solution and a subsequent resolution to the story.

Two picture stories, Hunting (Smith, 1982) and Picnic (Heaton, 1966), were further selected to represent the schematic sequential structure. These two stories have a clear macrostructure with an apparent time line in which things happen in a clear sequence of events. Both stories are considered schematic because there is a clear beginning, a well sign-posted sequence of events and an obvious ending to the stories.

Two picture stories, Walkman and Journey (Jones, 1980), were selected for the unstructured category since they meet all the conditions of the unstructured tasks. Both stories lack a clear time line and are based on an arbitrary sequence of events. In fact, the sequence of the events that happens in each story does not have an impact on the main theme of the story. As the pictures are loosely related to one another, it is possible to rearrange the pictures and create a new sequence of events without the main theme of the story being compromised. All six picture stories are included in Appendix 1.

8.3.4 Grounding

As grounding is the important variable in the design of the present study, the six picture stories discussed above are carefully selected to present either foreground or

foreground and background information. Three of the picture stories contain only foreground information and the other three contain background information as well as foreground information. While searching for suitable picture stories, I realized that most picture stories used in EFL teaching materials usually contain some background information. This background information is presented either explicitly or implicitly in the story. However, in the current study background information is to be carefully operationalized in the picture stories as it is hypothesized to influence the results. For this reason, finding picture stories that only contain foreground information has proved to be more difficult.

Three picture stories - Journey, Hunting and Football - are foreground narrative tasks. Journey and Football are principally foreground tasks with little or no background information. As the main story goes on, not much else occurs along the main events of these two picture stories. Nor is there any extra material presented to support or elaborate the events occurring in the foreground. However, in Hunting there is some implicit background information presented through the pictures of the story. This implicit background information is incorporated into the story when the main character starts thinking about or imagining probable events in the future. Although this was recognized as a potential shortcoming in the selection of Hunting as a foreground task, the Hunting task was eventually selected because all other alternative stories were not suitable for a variety of reasons. Therefore, Hunting is employed to represent a schematic sequential foreground task, with the awareness that its implicit background information might slightly influence the results.

Walkman, Picnic and Keys were selected as foreground and background tasks. In all these tasks, the main events happen in the foreground of the story and there are events that occur in the background as well. For Picnic and Keys, the events in the

background are incorporated into the events of the foreground and they impact on the story's outcome. However, in Walkman the many events that occur in the background have no impact on the main story of the foreground or on its outcome. This weak association between the events of foreground and background is clearly related to the fact that Walkman is an unstructured task. In fact, as the sequence of the events in Walkman is arbitrary, the events can happen in different orders. Hence, the background information does not relate to any specific event in the foreground. Nor does it impact on the story or its outcome. Table 8.2 summarizes the characteristics of the actual six tasks employed in Study Two showing the task characteristics that are being manipulated in the current study.

Table 8.2
Characteristics of the Oral Narrative Tasks in Study Two

Grounding	**Task Structure**		
	Unstructured	*Schematic structure*	*Problem-solution structure*
+ *Foreground* - *Background*	Journey	Hunting	Football
+ *Foreground* + *Background*	Walkman	Picnic	Keys

It is important to note that the notions of grounding and structure, as can be seen in the literature, require a high degree of interpretation. The interpretations made of the picture stories have been used in the pilot studies and they appeared to have worked in ways that we anticipated (See further discussions in section 8.3.5).

In order to avoid any results emerging from a practice effect, a counterbalanced design was considered for the administration of the tasks through data collection. As a result, all the participants performed three tasks but in different orders. Tables 8.3 and 8.4 show the counterbalanced sequence of tasks the participants will perform.

Table 8.3

Counterbalanced Sequence of Foreground Tasks

Sequence 1	Journey	Hunting	Football
Sequence 2	Hunting	Football	Journey
Sequence 3	Football	Journey	Hunting

Table 8.4

Counterbalanced Sequence of Foreground and Background Tasks

Sequence 1	Walkman	Picnic	Keys
Sequence 2	Picnic	Keys	Walkman
Sequence 3	Keys	Walkman	Picnic

8.3.5 Pilot Study

In order to find out whether the selected tasks were functioning in line with the theoretical assumptions of the study and to investigate whether there are other features in the picture stories which might intrude into language performance, all the selected tasks were piloted with 11 participants. The participants were adults aged between 19 and 35 and from a range of different countries such as Eritrea, Somalia, Turkey, Afghanistan and Iran. Three of the participants were Farsi speakers. All the participants were learning English as a second language at intermediate level in a college in London. The participants were informed of the purpose of the study and were assigned to either of the grounding categories to perform the three tasks. They were also asked to comment on the stories in terms of the clarity of the pictures and the understandability of the stories. The results of the performances of the first six participants indicated that one of the tasks, a Sempe picture story called 'The artist', seemed to be confusing to the participants. This picture story was substituted with another picture story, i.e. Hunting, with the same task structure and grounding

characteristics. The results also revealed that some pictures in one of the stories, Keys, were not clear because of the painting style of the story. To overcome this deficiency, an artist was employed to draw the Keys picture story in a lucid way. The copy of the Keys task available in Appendix 1 is, therefore, the drawn reproduction of the main story.

A second stage of the pilot study was carried out with another 5 participants performing the new set of the six picture stories and commenting on the clarity of them. The results of the second stage of the pilot study suggested that the selection of tasks was appropriate for the purpose of the study. It should be noted that a counterbalanced sequence of the tasks was also considered for the pilot study.

8.3.6 Participants in the Main Study

In the main study, the participants were 60 Iranian language learners studying English at Simin Educational Association in Tehran, Iran during the spring term 2003. They were all adult females aged between 18 and 34. They were studying English as a foreign language at an intermediate level and had been studying English in the same language school for at least 18 months. It should be noted that the participants in Study Two were different from those in Study One. The participants were Farsi speakers and had a similar language learning history both within the public schooling system and at the above-mentioned language school. However, they differed in terms of the period of time they had been studying English in the past, the contact they had with English outside the classroom and the purposes for which they were studying English.

As they had already taken part in similar testing situations in their language school and had performed similar tasks, they were all familiar with both the testing

conditions and the test format, i.e. oral narratives. The participants were randomly assigned to a foreground or a foreground-background group and one of the three counterbalanced sequences of tasks as mentioned in the previous sections. The counterbalanced design of the study with the two independent variables, the tasks and the number of the participants in each group are represented in Table 8.5.

Table 8.5

Design of Study Two

Grounding	***Task Structure: -struct., +struct., + struct.***	***Participants in Sequence 1***	***Participants in Sequence 2***	***Participants in Sequence 3***
Foreground	Journey, Hunting, Football	10	10	10
Foreground background	Walkman, picnic, Keys	10	10	10

8.3.7 Language Proficiency of the Participants

In Study One, all the participants were tested by an "Oxford Placement Test" before they took part in the study. This was initially carried out to make sure of the language proficiency level of the participants and to check the homogeneity of the groups. All the participants in both studies were placed in their levels based on their scores on a local institutional proficiency test. These tests are language proficiency tests, which assess listening, speaking, writing, reading skills and language use. Although they are developed locally their validity and reliability are regularly examined. Samples of these institutional tests can not be provided due to legal and administrative restrictions.

In Study One, the results of the Oxford Placement test were compared with the results of the institutional language proficiency test all the participants had taken two or three weeks before the study. As mentioned before, for practical reasons the grammar part

of the Oxford Placement Test was administered to the participants in Study One. However, all the participants had been examined on all different aspects of language ability through the institutional test. To make sure of the agreement between the two tests, a Pearson product-moment correlation was run. A relatively large coefficient correlation ($r = .56$) was observed between the two tests, suggesting that the local test was a reliable test of language proficiency to be used as the criterion in Study Two. Hence, the results of the local language test the participants had taken prior to Study Two were considered as an indicator of their language proficiency level. According to the criteria in the "Oxford Placement Test, participants who belong to band 4 are considered as intermediate level. Following the Oxford Placement criteria, in the current study all participants who had scored between 45 to 75 percent of their institutional test were considered intermediate proficiency level and were included in the study.

8.3.8 Setting of Administration

A parallel setting to Study One was arranged for the current study. As explained in Chapter V, all the participants were seen in a one-to-one setting with the researcher. The researcher met the participants individually and explained the purpose of the test to them. Each participant was randomly assigned to one of the grounding groups and to one of the three counterbalanced sequences of the tasks. Then, the instructions on how to perform the tasks were given to them. It is worth mentioning that in order to avoid any misunderstanding, all the explanations and instructions were given in the participants' native language, i.e. Farsi. Furthermore, the small language bits that appeared in some of the picture stories were translated to Farsi. The participants were given each of the picture stories at a time. They were told that they had 3 minutes to

look at each picture story and plan for what to say and how to narrate the story to the researcher who didn't have access to the stories. After that, they had 3 to 4 minutes to tell each story during which time their performance was recorded. The participants were also provided with some paper in case they wished to take notes but they were not forced to do so. They were reminded that they would not be allowed to use the notes while they were telling the stories. With each task, the participants looked at the pictures and told the story to the researcher who tape-recorded their performance. Then, the same procedures were followed for the second and third tasks one after the other. The notes participants made were all collected at the end of the test and kept for further analysis.

The rationale for using a 3-minute planning time is supported by the findings of a number of studies in the literature, which show that planning times of shorter than 3 minutes would not affect performance on tasks (Mehnert, 1998). At the same time it was felt that providing the participants with longer planning times is likely to contradict the practical requirements of the real-world testing situations.

8.4 Analytic Detailed Measures Adopted in Study Two

Following the procedures taken in Study One, and as analytic detailed measures are recommended in SLA research (See Chapters III and IV), a number of different fluency, accuracy and complexity measures are adopted to assess the participants' language performance on the six oral narrative tasks employed in this study. Detailed discussions about and definitions of different measures of fluency, accuracy and complexity are provided in Chapters IV and V. In the following section, a brief overview of the measures taken and explanations of some new measures adopted in Study Two will be presented.

8.4.1 Fluency Measures

As explained in Section 5.4.1, a wide range of different fluency measures was employed in Study One. To measure repair fluency, the number of false starts, reformulations, repetitions and replacement were recorded. Measures of the total amount of silence, mean number of pauses and mean length of pauses were taken to represent temporal aspects or breakdown fluency. Speech rate, length of run and proportion of time spoken were other measures of temporal fluency that were taken as an indication of how fast, how linguistically dense or how lengthy each performance was. Definitions of all these measures are provided in Chapter V, Section 5.4.1.

In addition to the measures adopted in Study One, some further measures of fluency were considered for Study Two. In Study Two the measures of breakdown fluency, as termed by Skehan (2003), were investigated in terms of the breakdown happening in the middle of clauses as compared with the breakdown occurring at clause boundaries. In fact, as mentioned while discussing the results of Study One, i.e. Section 7.4, Study two attempted to investigate whether the mid-clause or the end-clause pauses would make the language performance less fluent. In doing so, it was necessary that the pauses of longer than .4 a second were identified and recorded separately as mid-clause pauses contrasted with end-clause pauses. As a result, the number of mid-clause and end-clause pauses, the length of mid-clause and end-clause pauses and the total amount of mid-clause and end-clause silence would be the additional measures of fluency employed in Study Two.

8.4.2 Complexity Measure

Parallel to Study One, the data are coded for AS units that contain independent clauses, subordinate clauses and sub-clausal units. In fact, in order to compare the

results of the complexity measure from both studies, it is necessary to consider the same complexity measure as employed in Study Two, i.e. an index of subordination in AS units. Definitions of AS unit, subordinate clauses and sub-clausal units as well as the rationale for measuring complexity through subordination in AS unit of analysis are provided in detail in Chapters IV and V.

In addition to the index of subordination, an extra measure of complexity was required to provide this research with a broader perspective to the issue of complexity of language performance on tasks. A number of researchers have mentioned lexical variety as an indicator of language learners' active vocabulary and an aspect of language complexity (Mehnert, 1998; Robinson, 2001). Different measures of lexical variety are used in a wide range of educational and linguistic research. These measures usually reflect the variety of active vocabulary employed by a speaker or writer and are typically measured by the Type-Token Ratio (TTR): the ratio of different words (types) to the total number of words (tokens) used (See malvern & Richards, 2002). It is worth noting that TTR has proved to be problematic because it is a function of sample size, i.e. larger samples of words will give a lower TTR. To overcome this problem, some researchers have developed new measures of lexical variety (Richards, 1987). Recently, Malvern and Richards (2002) have used an innovative measure, D, which is based on mathematically modelling how new words are introduced to larger and larger language samples. The authors claim that D is a more appropriate measure because it is independent of sample size and allows valid comparisons between speakers who produce varying quantities of linguistic data. Malvern and Richards (2002) have also produced a software, *vocd*, to calculate lexical variety, which is available through the CLAN program in the CHILDES project (www.childes.psy.cmu.edu). Hence, in the current study *vocd* measure of lexical

variety was employed as an extra measure of complexity so that the language complexity of learners' performance can be investigated in a wider context.

8.4.3 Accuracy Measure

Following procedures taken in Study One, an index of error-free clauses is employed to represent the accuracy of each performance. The detailed discussions of different accuracy measures and the rationale for adopting a general measure of accuracy in this research are provided in Chapters IV and V.

8.5 Data

The audio recordings of the performance of all 60 participants of Study Two were digitized using the Goldwave software (See Chapter V for a description of the Goldwave software). The digitized data were recorded onto CDs so that they are compatible with the use of different computer software. Then, using the SoundScriber[2] software, the whole data were transcribed. The data were then word-processed. Using Goldwave, all the pauses longer than .4 of a second were measured and marked in the data. The length of each performance was further digitally measured and inserted in the transcribed data. Parallel to Study One, all the data were coded for a wide range of fluency, accuracy and complexity measures (See Appendix 4 for all the coding symbols and Appendix 5 for samples of the coded data). The coding procedures were carried out in a similar way to Study One. The only difference between the coding systems of the two studies is the addition of a new set of pauses, i.e. mid-clause versus end-clause pauses, in Study Two.

---

[2]SoundScriber is a program for Windows which aids in transcription of digitized sound files and is available as freeware online at www.lsa.umich.edu/eli/micase

8.5.1 Coding the Data and Inter-Rater Reliability

Similar to the procedures taken in Study One, 10% of the data was coded by an experienced researcher, against which all the coded data are tested. Table 8.6 demonstrates the inter-rater reliability coefficient of the coded data.

Table 8.6

Inter-Rater Reliability Coefficient of the Coded Data

Measures	*AS unit*	*Dependent Clauses*	*Error-free clause*	*Reformulation*	*Replacement*	*Repetition*	*False start*
Coefficient	.95	.91	.96	.99	1.00	1.00	.99

As indicated in Table 8.6, a reliability coefficient of higher than 90% was achieved for coding the data on all measures of accuracy, fluency and complexity. It is worth mentioning that the temporal fluency measures, e.g. the mean length of pauses and mean number of pauses, were not included in the inter-rater reliability test since they were all digitally measured and the chances of obtaining errors seemed to be minimized.

8.5.2 Computer Programs Used to Analyze the Data

As partly explained in section 8.5, a number of computerized programs and software were used to analyze the data. SoundScriber was used to transcribe the audio-recorded performances. Goldwave Audio Editor was employed for digitizing the tape-recorded performances and for measuring the pauses and the amount of speaking time for each performance. As mentioned in Section 5.5.2, in order to read the coded transcripts of the data and to calculate different coded measures in the transcripts, a special computer program was used. This program was primarily designed and developed for analyzing language performance data. However, to meet the specific requirements of the data analyses in the current study, the program was adapted at

later stages to deal with the data more efficiently, i.e. to distinguish between mid-clause and end-clause pauses. The program is able to understand the codings and to calculate how many of each individual measure were found in the performance of each participant on each task. Once the program calculated the figures for each measure, or calculated the ratio required for some other measures, a SPSS data set was made and the figures were transferred to the data set. SPSS 9.0 was used to test the data for a number of various statistical analyses, which will be discussed in the chapter that follows.

CHAPTER IX

Analyses and Results: Study Two

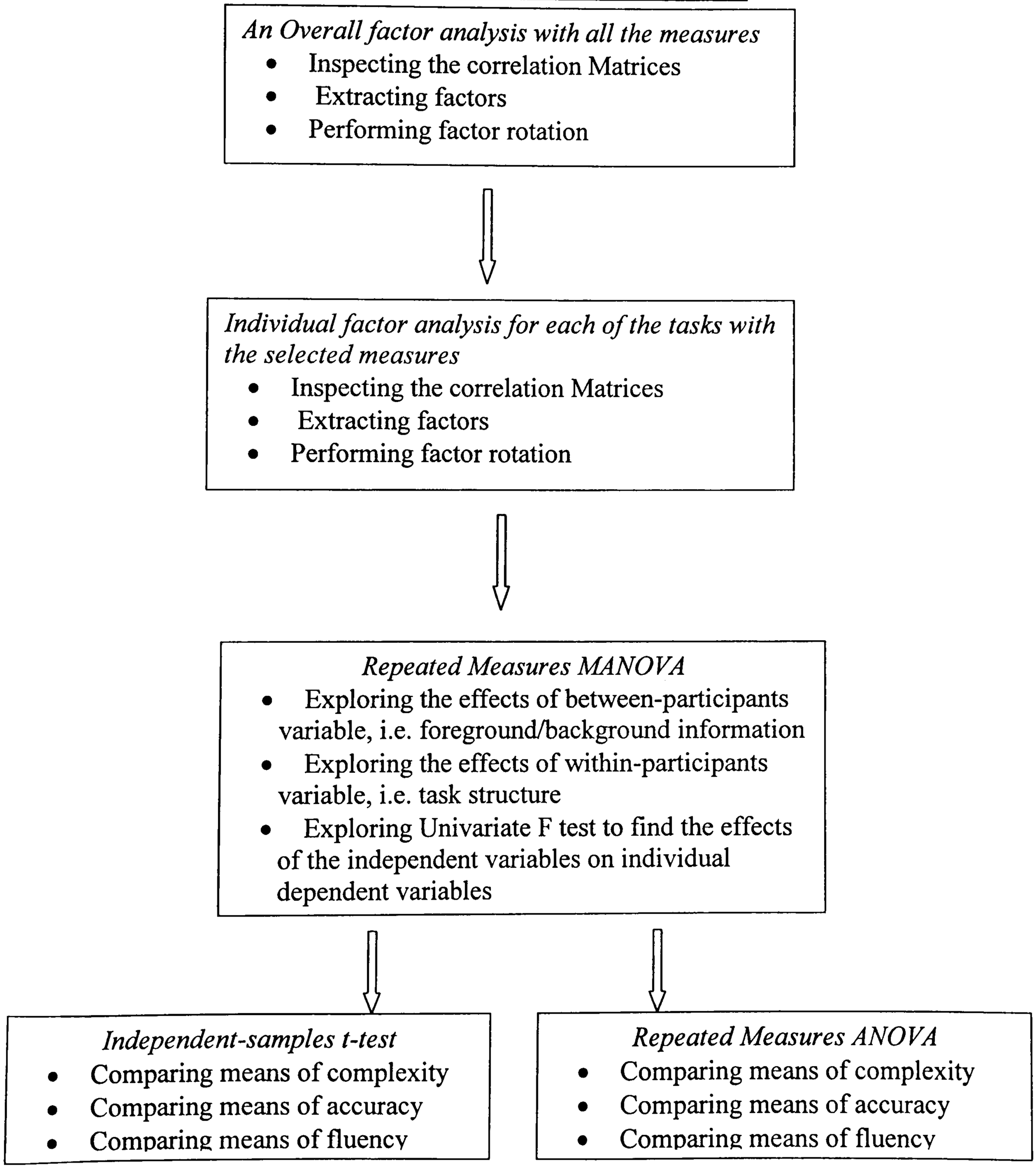
9.1 Introduction

As discussed in Chapter VIII, a 2 x 3 factorial design was selected for the current study with the task structure and grounding, i.e. foreground and background information, as the two independent variables. Task structure was taken as the within-participants variable indicating three degrees of structure, i.e. unstructured, schematic structured and problem-solution structured. Grounding was taken as the between-participant variable and was represented through two levels of information available in the narrative tasks, i.e. tasks with foreground information and tasks with both foreground and background information. The dependent variables of the study were fluency, accuracy and complexity of the performances. In order to test the three hypotheses of this study, a number of various statistical analyses were employed.

Following the procedures taken in Study One, that is, to uncover whether the different dependent variables of the study truly delineate separate factors, a series of factor analyses were conducted for each individual task. Therefore, the first statistical analysis employed in this study was a series of factor analyses. Based on the results of the factor analyses, a repeated measures MANOVA was performed to find out whether the independent variables had any significant effect on the language performance of the participants. Once the results of the MANOVA showed that the participants' performance was influenced by the independent variables, it was

justifiable to continue the analyses with further statistical procedures. In so doing, a series of t-tests were conducted to investigate the detailed effects of grounding on different dependent variables. In order to explore the effects of task structure on different aspects of performance and to compare these results with the findings of Study One as well as the previous research, a series of ANOVAs were performed. A flow chart of these statistical procedures is presented in Figure 9.1.

Figure 9.1
A Flow Chart of the Statistical Procedures Used in Study 2



9.2 Statistical Analyses

9.2.1 Underlying Factors in Language Performance

As mentioned in the previous chapter, a number of various dependent variables were initially employed during the coding and analysis of the data for the current study. As a further development to Study One, measures of mid-clause and end-clause pauses as well as the percentage of accuracy for clauses of different length were employed along with all the measures used in Study One. However, for both practical and theoretical reasons, including all these measures in factor analysis or multivariate analysis was inappropriate.

In assessing the suitability of the data for factor analysis, two principle issues of strength of relationship among the variables and sample size were initially considered. With regard to the strength of the relationship, the correlation matrices were inspected and the results will be discussed in a following section. Regarding sample size, although larger samples of about 300 participants are assumed to produce more reliable results, smaller sample sizes are accepted if there are strong reliable correlations and a few distinct factors (Tabachnick & Fidell, 1996). Moreover, recent researchers (Stevens, 1996; Nunnally, 1978) suggest that it is not the overall sample size that is of a concern, rather it is the ratio of the subjects to the items, i.e. the dependent variables. Tabachnic and Fidell (1996) recommend that a ratio of 5 cases for each variable to be factor analysed is adequate in most research studies. Given that 30 participants performed each task in the present study, a factor analysis that includes 6 items was thus justifiable. Therefore, in order to meet the assumptions of factor analyses, the best representatives for each of the fluency, accuracy and complexity measures had to be selected. In so doing, a collective factor analysis was first conducted with all the fluency measures for each of the tasks. These primary

factor analyses showed that, from among the fluency measures, false start, speech rate, number of mid-clause pauses and number of end-clause pauses had high loadings and were stable in their loading factors across all six tasks (See Appendix 6 for the details of these factor analyses). Consequently, false start was selected to represent repair fluency measures, speech rate to designate speed fluency and number of pauses to stand for break down fluency measures. For the accuracy measure, the overall ratio of error free clauses was included in the analyses. Given that Study Two was based on Study One and primarily designed to explore the nature of complexity, the same measure of complexity, i.e. ratio of subordination was used to create an analogous situation between the two studies. Therefore, this measure of complexity was included in the items of factor analysis. Finally, the six selected measures were utilized and an individual factor analysis was conducted for each of the tasks.

Prior to performing the analyses, the suitability of the data for factor analysis was investigated. First, as explained in the previous paragraph, the issue of sample size was taken into consideration and a selection of 6 dependent variables was justified to provide a clear picture for each factor analysis. Second, inspections of the correlation matrices revealed the presence of many coefficients of .3 and above for all the tasks, confirming the suitability of the data for factor analysis. In addition, the Kaiser-Meyer-Oklin values were above .60 for all the tasks, and were practically equal to or above the recommended value of .60 (Kaiser, 1974). Barlett's Test of Sphericity reached statistical significance, supporting the factorability of each correlation matrix.

The results shown in Tables 9.1 to 9.6 demonstrate all factor loadings of .40 and above, and the communality scores of the measures which indicate the amount of variance that is explained by the factor. It is worth mentioning that in contrast with Study One, a unified number of factors or a fixed set of measures was not extracted

from the analyses for all the tasks. It can be argued that the main influence on obtaining such different results is the low number of participants in this study, as compared with a higher number of participants doing each of the tasks in Study One. Discussions of the factors and loadings of the measures for the six tasks will follow.

Table: 9.1

Factor Analysis for the Journey Task

Measures	Factor 1	Factor 2	Communality
False start		.843	.773
Accuracy	-.517		.323
Complexity		.593	.416
Speech rate	-.705		.649
Mid-clause pauses	.841		.751
End-clause pauses	.834		.705

Table: 9.2

Factor Analysis for the Hunting Task

Measures	Factor 1	Factor 2	Communality
False start		.840	.719
Accuracy	.773		.612
Complexity		-.548	.304
Speech rate	.873		.784
Mid-clause pauses	-.777	.464	.819
End-clause pauses	-.642		.564

Table: 9.3

Factor Analysis for the Football Task

Measures	Factor 1	Factor 2	Factor 3	Communality
False start		.842		.730
Accuracy			.907	.846
Complexity		.762		.592
Speech rate	-.846			.722
Mid-clause pauses	.889			.804
End-clause pauses		-.431	.622	.637

Table: 9.4

Factor Analysis for the Walkman Task

Measures	Factor 1	Factor 2	Communality
False start	-.447	.655	.629
Accuracy	.691		.478
Complexity	.781		.618
Speech rate	.775		.695
Mid-clause pauses	-.386	.824	.828
End-clause pauses		.755	.589

Table: 9.5

Factor Analysis for the Picnic Task

Measures	Factor 1	Communality
False start	-.703	.494
Accuracy	.780	.608
Complexity	.588	.345
Speech rate	.848	.719
Mid-clause pauses	-.813	.661
End-clause pauses	-.614	.377

Table: 9.6

Factor Analysis for the Keys Task

Measures	Factor 1	Factor 2	Communality
False start	.394		.155
Accuracy		.861	.745
Complexity		.854	.733
Speech rate	.831		.772
Mid-clause pauses	.926		.858
End-clause pauses	.666		.513

As the results of the factor analyses reveal, there are certain differences across the tasks in terms of the number of factors extracted, the loadings of the measures on the factors and the order in which the extracted factors appear. For instance, the number of factors extracted for four of the tasks - Journey, Hunting, Walkman and Keys - are two. But for Picnic a single factor is extracted and for Football three factors are extracted. Accuracy and complexity have loaded on the same factor for tasks containing both foreground and background information but on different factors for tasks which contain only foreground information. Nonetheless, despite such apparent disparity, there is a good deal of meaningful congruity across the tasks if they are compared with each other in a relevant context in terms of the independent variables of the study.

As defined in the previous chapter, to investigate the effect of grounding on syntactic complexity of second language performance, the tasks are primarily selected based on the foreground or foreground and background information they bear. Therefore, it is appropriate to investigate and compare the factor analyses within their grounding

groups. For foreground information tasks, the results of the factor analyses show that accuracy and complexity have always loaded on two different factors, suggesting that the two variables refer to two marginally correlated or perhaps two distinct constructs. Complexity and false start always loaded highly on Factor 2 suggesting that for these tasks there is a correlation between complexity and false starts.

Speech rate and number of mid-clause pauses have loaded on Factor 1 suggesting that the speed with which participants perform the tasks relates to the number of times they pause in the middle of a clause. The association between speed of performance and the number of pauses in a performance is predictable, as the more the number of pauses would inevitably lead to the performance being slower. However, the results of this factor analysis show that a stronger association exists between the speech rate and the number of mid-clause pauses rather than the end-clause pauses. This factor also contains accuracy for Journey and Hunting, implying that the faster the speed and the fewer the number of pauses, the more accurate the language performance has been in foreground tasks. For Football, accuracy and number of end-clause pauses load on a third factor, implying that the more accurate performances included a fewer number of end-of-clause pauses.

For tasks that contain both foreground and background information, the results show that Walkman and Keys have loaded on two factors, while Picnic has loaded on a single factor. Strikingly, for all these tasks, accuracy and complexity have loaded on one factor, suggesting that the more complex the performance in tasks, the more accurate they will be. In effect, it suggests that there is an association between complexity and accuracy for all the tasks with foreground and background information. For Picnic, all six measures of accuracy, fluency and complexity have loaded on one single factor, demonstrating a high degree of go-togetherness among

different aspects of performance in this particular task. For Walkman, the number of mid and end-clause pauses and false start load on Factor 2, while speech rate loads along with accuracy and complexity on Factor 1. For Keys, all different measures of fluency load on Factor 1 and accuracy and complexity load on Factor 2.

As the results clearly demonstrate, grounding plays a significant role on the grouping of the dependent variables in the factor analyses. When background information is incorporated into the tasks, a high amount of integrity is observed between accuracy and complexity. In such an instance, the form construct, i.e. complexity and accuracy measures, distinguishes itself from the meaning construct, i.e. fluency measures. Whereas, when the task lacks background information, the immediate association of complexity and accuracy disappears and the form construct loads on two different factors, each being associated with a number of fluency measures. A detailed discussion of the relationship between the dependent variables will be presented in Chapter X. The correlation matrices for all the tasks are shown in Tables 9.7 to 9.13.

Table: 9.7

Correlation Matrix for the Journey Task

correlations	J. FALST	J. ACCU	J. COMP	J. SPRAT	J. NOPS1	J. NOPS2
J .FALSTART	1.00	-.015	.163	.193	.295	.208
J .ACCURAC	-.015	1.00	.150	.281	-.146	-.391
J .COMPLEX	.163	.150	1.00	.186	-.089	-.165
J .SPCHRATE	.193	.281	.186	1.00	-.569	-.388
J. NOPAUSE1	.295	-.146	-.089	-.569	1.00	.598
J .NOPAUSE2	.208	-.391	-.156	-.388	.598	1.00

Table: 9.8

Correlation Matrix for the Hunting Task

correlations	H. FALS	H. ACCU	H. COMP	H. SPRAT	H. NOPS1	H. NOPS2
H .FALSTART	1.00	.013	-.153	-.039	.484	.232
H .ACCURAC	.013	1.00	.137	.570	-.403	-.346
H .COMPLEX	-.153	.137	1.00	-.087	-.062	-.112
H .SPCHRATE	-.039	.570	-.087	1.00	-.611	-.351
H. NOPAUSE1	.484	-.403	-.062	-.611	1.00	.616
H. NOPAUSE2	.232	-.346	-.112	-.351	.616	1.00

Table: 9.9

Correlation Matrix for the Football Task

correlations	F. FALS	F. ACCU	F. COMP	F. SPRAT	F. NOPS1	F. NOPS2
F .FALSTART	1.00	.015	.363	.017	.200	-.266
F .ACCURAC	.015	1.00	-.064	.012	-.088	.258
F .COMPLEX	.363	-.064	1.00	.077	-.009	-.231
F. SPCHRATE	.017	.012	.077	1.00	-.541	-.131
F. NOPAUSE1	.200	-.088	-.009	-.541	1.00	.167
F. NOPAUSE2	-.266	.258	-.231	-.131	.167	1.00

Table: 9.10

Correlation Matrix for the Walkman Task

correlations	W. FALS	W. ACCU	W. COMP	W. SPRAT	W. NOPS1	W. NOPS2
W. FALSTART	1.00	-.257	-.254	-.501	.711	.153
W. ACCURAC	-.257	1.00	.371	.351	-.243	-.119
W. COMPLEX	-.254	.371	1.00	.538	-.333	-.135
W. SPCHRATE	-.501	.351	.538	1.00	-.522	-.095
W. NOPAUSE1	.711	-.243	-.333	-.522	1.00	.402
W. NOPAUSE2	.153	-.119	-.135	-.095	.402	1.00

Table: 9.11

Correlation Matrix for the Picnic Task

correlations	P. FALS	P. ACCU	P. COMP	P. SPRAT	P. NOPS1	P. NOPS2
P. FALSTART	1.00	-.496	-.268	-.549	.463	.265
P. ACCURAC	-.496	1.00	.356	.664	-.542	-.255
P. COMPLEX	-.268	.356	1.00	.378	-.330	-.391
P. SPCHRATE	-.549	.664	.378	1.00	-.627	-.377
P. NOPAUSE1	.463	-.542	-.330	-.627	1.00	.516
P. NOPAUSE2	.265	-.255	-.391	-.377	.516	1.00

Table: 9.12

Correlation Matrix for the Keys Task

correlations	K. FALS	K. ACCU	K. COMP	K. SPRAT	K. NOPS1	K. NOPS2
K. FALSTART	1.00	-.084	.053	-.074	.351	.113
K. ACCURAC	-.084	1.00	.521	.119	.41	.207
K. COMPLEX	.053	.521	1.00	.251	.006	.056
K. SPCHRATE	-.074	.119	.251	1.00	-.764	-.349
K. NOPAUSE1	.351	.041	.006	-.764	1.00	.441
K. NOPAUSE2	.113	.207	.056	-.349	.441	1.00

Before moving to any further discussion of the statistical analyses, I would like to discuss the inconsistency of the results of the factor analyses in Study 2 as compared to those obtained from Study One. In Study One, there were 4 tasks and 80 participants who performed each of the four tasks. In other words, there were 80 performances for each single task. In Study Two, however, there were six tasks and the 60 participants performed the tasks in a between-participant design. This means that only 30 participants performed each of the tasks in Study Two. Therefore, the inconsistency between the results of the factor analyses could partly be explained by the fact that the smaller number of the participants in Study Two could have influenced the results of the factors (Tabachnic & Fidell, 1996).

9.2.2 MANOVA: Effects of the Independent Variables

A 2 x 3 repeated measures MANOVA was performed to investigate the effects of task structure and grounding on different dependent variables of the study. A MANOVA is used because there are three dependent variables, i.e. accuracy, complexity and fluency, and two independent variables, i.e. task structure and grounding in the study. As grounding is a between-participants and task structure a within-participants variable, a mixed between-within MANOVA was used. Meanwhile, a repeated measures was employed because the means to be tested were derived from the same participants measured on three different tasks.

Based on the results of the initial factor analyses explained in the previous section, six dependent variables of false start, accuracy, complexity, speech rate, number of mid-clause pauses and number of end-clause pauses were selected and used in the MANOVA. The criterion for selecting these measures was the consistency and high loadings of the measures across the tasks. Prior to performing the MANOVA

analysis, the suitability of the data for such analysis was tested. Different tests were conducted to make sure that the data were of the right sample size and appropriate normality and linearity. As MANOVA is quite sensitive to outliers, Univariate and multivariate tests of outliers were also run. Results of the repeated measures MANOVA are presented in Table 9.13.

The results of the analysis clearly indicate that the dependent variables are influenced by both task structure and grounding. As regards the between-participants variable, the analysis reveals that there is a statistically significant difference between the foreground and foreground and background tasks (*Pillai's* = .474, $F = 7.95$, $P = .001$, $\eta^2 = .47$). Regarding the within-participants variable, a significant difference is observed across the tasks in terms of task structure (*Pillai's* = .484, $F = 5.95$, $P = .001$, $\eta^2 = .24$) with the differences being concentrated on five of the six measures used in MANOVA. Furthermore, the interaction between task and grounding proved to be significant for one of the dependent variables (*Pillai's* = .232, $F = 2.45$, $P = .008$, $\eta^2 = .11$).

Table 9.13

Results of Repeated Measures MANOVA

Between-Participants Effect

Effects	**Pillai's Value**	***F***	***BGdf***	***WGdf***	**Sig.**	**Eta Sq**
Grounding	.474	7.95	6	53	.001*	.47

* Significant differences are reached.

Within-Participants Effect

Effects	**Pillai's Value**	***F***	***BGdf***	***WGdf***	**Sig.**	**Eta Sq**
Task	.484	5.95	12	224	.001*	.24
Task x Grounding	.232	2.45	12	224	.008*	.11

* Significant differences are reached.

The measure of effect size is continuously considered in the current study since the reliability of significant findings is questioned if adequate information of the effect size is not provided (Fulcher & Marquez Reiter, 2003). As explained in Chapter VII, effect size describes the amount of the total variance in the dependent variable that is predictable from knowledge of the levels of the independent variable. As the results indicate the two effect size values emerging from the independent variables are noticeable. Furthermore, the value of the effect size for grounding is larger than the value of the effect size for task structure, indicating that a larger amount of the variance in the participants' performance is explained by the influence of the grounding variable.

So far the results of the MANOVA have clearly indicated that both grounding and task structure have influenced the dependent variables. However, in order to find out which dependent variables have been significantly influenced, a Univariate F test is required. When the results for the dependent variables were considered separately through the Univariate F test, using a Bonferoni adjusted alpha level[1] (recommended by Tabachnick & Fidell, 1996), significance was reached for five measures of accuracy, complexity, false start, number of mid-clause pauses and number of end-clause pauses. Significance was further observed for the interaction between task and grounding for the false start measure. Table 9.14 summarizes the results of the Univariate F test of within-participant effects.

---

[1]A Bonferoni adjustment to alpha level is usually adopted in order to prevent an inflated risk of Type 1 errors. In fact, a more stringent level is being set to avoid rejecting a null hypothesis when it is true.

Table: 9.14

Univariate Test of Within-Participant Effect

Source	Measure	Sum of squares	df	Mean square	F	Sig.	Eta. Sq.
Task Structure	Accuracy	.469	2	.235	13.28	.001*	.186
	Complexity	1.37	2	.685	15.94	.001*	.216
	False start	36.63	2	18.31	3.08	.05*	.051
	No. of pauses mid	89.90	2	44.27	13.13	.001*	.185
	No. of pauses end	86.68	2	43.43	7.20	.003*	.110
	Speech rate	83.171	2	41.58	.37	.69	.006
Task Structure x Grounding	Accuracy	.012	2	.006	.035	.966	.001
	Complexity	.155	2	.77	1.80	.169	.03
	False start	101.34	2	50.67	8.54	.003*	.128
	No. of pauses mid	175.54	2	87.77	2.57	.08	.043
	No. of pauses end	47.67	2	23.83	3.96	.06	.064
	Speech rate	236.5	2	118.25	1.06	.35	.018

* Significant differences are reached.

The results show that there are significant differences across the tasks for five of the six dependent variables. However, more specific comparisons are required to find out where the differences are located. Therefore, pairwise comparisons are required to demonstrate where the statistically significant differences are located for each dependent variable across the tasks. Tables 9.15 to 9.20 show the results of the pairwise comparisons across the tasks.

Table: 9.15

Pairwise Comparisons between Tasks: Accuracy

Tasks	Journey/Walkman	Hunting/Picnic	Football/Keys
Journey/Walkman	-	.001	.001
Hunting/Picnic		-	NS
Football/Keys			-

Table: 9.16

Pairwise Comparisons between Tasks: Complexity

Tasks	Journey/Walkman	Hunting/Picnic	Football/Keys
Journey/Walkman	-	.001	.001
Hunting/Picnic		-	NS
Football/Keys			-

Table: 9.17

Pairwise Comparisons between Tasks: False Start

Tasks	Journey/Walkman	Hunting/Picnic	Football/Keys
Journey/Walkman	-	.03	NS
Hunting/Picnic		-	NS
Football/Keys			-

Table: 9.18

Pairwise Comparisons between Tasks: Speech Rate

Tasks	Journey/Walkman	Hunting/Picnic	Football/Keys
Journey/Walkman	-	NS	NS
Hunting/Picnic		-	NS
Football/Keys			-

Table: 9.19

Pairwise Comparisons between Tasks: No. of Mid-Clause Pauses

Tasks	Journey/Walkman	Hunting/Picnic	Football/Keys
Journey/Walkman	-	.001	.001
Hunting/Picnic		-	NS
Football/Keys			-

Table: 9.20

Pairwise Comparisons between Tasks: No. of End-Clause Pauses

Tasks	Journey/Walkman	Hunting/Picnic	Football/Keys
Journey/Walkman	-	NS	.001
Hunting/Picnic		-	.001
Football/Keys			-

For the three measures of accuracy, complexity and number of mid-clause pauses, the unstructured tasks are significantly different from both structured tasks, but the structured tasks are not different from one another. In fact, the performance in the unstructured task is significantly less accurate (means of accuracy: Journey/Walkman = .30, Hunting/Picnic = .41, Football/Keys = .41), less complex (means of complexity: Journey/Walkman = 1.30, Hunting/Picnic = 1.51, Football/Keys = 1.41) and less fluent on number of mid-clause pauses (means of number of mid-clause pauses: Journey/Walkman = 16.35, Hunting/Picnic = 12.2, Football/Keys = 11.2). For false start, the unstructured task is different from the schematic structured tasks but not from the problem-solution structured tasks (means of false start: Journey/Walkman = 4.9, Hunting/Picnic = 3.88, Football/Keys = 4.01). In other words, performance in the schematic structured tasks, i.e. Picnic and Hunting, is significantly different from the performance in the unstructured tasks. For the number of end-clause pauses, the problem-solution tasks are significantly more fluent than the other two groups of tasks (means of number of end-clause pauses: Journey/Walkman = 6.83, Hunting/Picnic = 6.28, Football/Keys = 5.16). For the measure of speech rate no significant difference has been observed across the tasks (means for speech rate: Journey/Walkman = 99.19, Hunting/Picnic = 100.79, Football/Keys = 100.40). The results of the repeated measures MANOVA clearly justify the continuation of the analyses with further detailed measures for both the dependent variables of the study. The details of where the significant differences are located for the between-participant (grounding) and within-participant (task structure) variables will follow in the next sections.

9.2.3 T-Tests: Effects of Grounding

In order to investigate the effects of grounding on dependent variables of the study, a series of t-tests were conducted for all measures of fluency, accuracy and complexity. Independent-samples t-tests were used to compare mean scores of the participants' performances on tasks that contained foreground information with the mean scores of tasks which contained foreground and background information, in terms of their fluency, accuracy and complexity. Through the t-tests, pairs of tasks that belonged to the same type of task structure but differed in the type of grounding were statistically compared. In other words, t-tests have been performed to compare Journey with Walkman, Hunting with Picnic, and Football with Keys. The results of the t-tests including the t values, the significance levels, means and standard deviations for all the tasks and an indication of whether significance was reached are demonstrated in Tables 9.21a, 9.21b and 9.21c. Table 9.21a will demonstrate the results of t-test comparing means of the different measures of Journey and Walkman tasks; Table 9.21b will show the results of the t-tests comparing means of the various measures of Hunting and Picnic; and Table 9.21c will present the results of the t-tests comparing means of the different measures of the Football and Keys tasks. *Vocd* measure of lexical variety has been excluded from the further statistical analyses since the results obtained for this measure were very unclear and thus difficult to interpret. In any case, since the main focus of the analysis is concerned with syntactic structural complexity, and *vocd* is a measure of lexical variety, a decision was made to jettison this supplementary measure, because it would not have added directly to the analyses. The discussions of the results of the t-tests will be presented in section 9.3.1 later in the current chapter.

PAGE NUMBERING AS ORIGINAL

Table: 9.21a

Results of T-tests: Effects of Grounding for Journey vs. Walkman

	T	*P*	Foreground Means for Journey	Foreground + background Means for Walkman
Reformulation	.87	.38	1.00 (1.17)	1.33 (1.72)
False start	2.56	.01*	3.56 (2.89)	6.23 (4.90)
Replacement	3.44	.001*	1.13 (1.15)	2.86 (2.31)
Repetition	2.57	.01*	2.56 (2.67)	5.46 (5.56)
Accuracy	.34	.73	.31 (.17)	.30 (.16)
Complexity	2.01	.04*	1.24 (.17)	1.36 (.27)
Length of run	.29	.77	4.22 (1.91)	4.36 (1.71)
Speech rate	.80	.42	101.54 (22.21)	96.84 (23.17)
No. of pauses mid-clause	1.79	.07	13.76 (8)	18.93 (13.4)
No. of pauses end-clause	3.78	.001*	4.86 (2.8)	8.80 (4.95)
Total silence mid-clause	1.77	.08	11.31 (9.1)	16.79 (14.23)
Total silence end-clause	2.76	.008*	3.8 (2.8)	6.59 (4.72)
Prop. time spoken	1.19	.23	.83 (.08)	.86 (.07)
Pause length mid-clause	.92	.36	.75 (.22)	.81 (.27)
Pause length end-clause	.25	.79	.75 (.36)	.73 (.24)

Table: 9.21b

Results of T-tests: Effects of Grounding for Hunting vs. Picnic

	T	*P*	Foreground Means for Hunting	Foreground + background Means for Picnic
Reformulation	.12	.89	.86 (1.19)	.83 (.79)
False start	.76	.44	4.13 (2.30)	3.63 (2.73)
Replacement	.77	.44	2.23 (1.67)	1.9 (1.66)
Repetition	2.34	.02*	2.6 (2.26)	4.63 (4.16)
Accuracy	.06	.95	.42 (.16)	.41 (.21)
Complexity	2.16	.03*	1.43 (.16)	1.59 (.38)
Length of run	.63	.53	4.19 (2.37)	4.58 (2.33)
Speech rate	.53	.59	102.51 (24.44)	99.06 (25.68)
No. of pauses mid-clause	.27	.78	11.9 (6.79)	12.5 (9.8)
No. of pauses end-clause	1.80	.07	5.40 (2.28)	7.16 (4.85)
Total silence mid-clause	1.15	.25	9.68 (7.2)	13.15 (14.77)
Total silence end-clause	2.02	.04*	3.54 (2.5)	5.68 (5.19)
Prop. time spoken	.05	.95	.86 (.07)	.86 (.10)
Pause length mid-clause	1.24	.22	.75 (.25)	.98 (.95)
Pause length end-clause	1.66	.10	.61 (.22)	.7 (.23)

Table: 9.21c

Results of T-tests: Effects of Grounding for Football vs. Keys

	T	*P*	Foreground Means for Football	Foreground + background Means for Keys
Reformulation	1.84	.07	.63 (.61)	1.16 (1.46)
False start	3.72	.001*	2.66 (2.05)	5.36 (3.39)
Replacement	3.17	.002*	1.26 (1.28)	2.6 (1.9)
Repetition	2.16	.03*	2.7 (3.08)	4.93 (4.74)
Accuracy	.16	.87	.42 (.16)	.41 (.14)
Complexity	4.77	.001*	1.28 (.16)	1.54 (.24)
Length of run	.92	.35	4.86 (3.16)	4.24 (1.5)
Speech rate	1.43	.15	104.81 (24.56)	95.99 (22.91)
No. of pauses mid-clause	2.57	.01*	9 (5.5)	13 (.23)
No. of pauses end-clause	2.07	.04*	4.3 (3.08)	6.03 (3.38)
Total silence mid-clause	2.44	.01*	7.86 (5.9)	12.64 (8.9)
Total silence end-clause	2.45	.01*	2.58 (1.94)	4.46 (3.70)
Prop. time spoken	.41	.67	.86 (.07)	.86 (.07)
Pause length mid-clause	1.14	.25	.82 (.31)	.94 (.47)
Pause length end-clause	1.17	.24	.62 (.23)	.71 (.33)

* Significant differences are reached.

9.2.4 ANOVAs: Effects of Task Structure

As a reminder, it should be mentioned that task structure is the within-participant variable of the study with three different levels of unstructured, schematic sequential and problem-solution structure. This means that each participant performed three different tasks with different degrees of structure. Therefore, the statistical analysis required to investigate the effects of task structure on different dependent variables of the study was repeated measures ANOVA. A series of repeated measures ANOVAs were run on all various measures of fluency, accuracy and complexity. Where significance was reached a planned comparison was run to explore where the significant differences were located. To avoid any inflated Type 1 error, the Bonferoni adjusted level has been considered.

It should be noted that, as task structure was the within-participant and grounding the between-participant variable, 50% of the participants performed three tasks of foreground information and 50% performed three tasks of both foreground and background information. As a result, a series of ANOVAs was required to compare performance in the three foreground information tasks, Journey, Hunting and Football. A second series of ANOVAs were employed to compare the three foreground and background information tasks, Walkman, Picnic and Keys. The results of the ANOVAs, F values, significant levels, means and standard deviations, effect size and an indication of where differences reached significance for each of the measures are shown in Table 9.22 for the foreground information tasks and in Table 9.23 for the foreground and background information tasks.

Table: 9.22

Results of the ANOVAs for Journey, Hunting and Football

Measures	*F*	*P*	Task / Structure: Journey	Hunting	Football	Sig. differences	Eta Squ.
Reformulation	1.16	.32	1.00 (SD= 1.17)	.86 (SD= 1.19)	.63 (SD= .61)		.07
False start	5.20	.03*	3.56 (SD= 2.89)	4.13 (SD= 2.30)	2.66 (SD= 2.05)	F vs. J H	.27
Replacement	4.83	.03*	1.13 (SD= 1.5)	2.23 (SD= 1.67)	1.26 (SD= 1.28)	H vs. J F	.23
Repetition	.04	.95	2.56 (SD= 2.67)	2.60 (SD= 2.26)	2.70 (SD= 3.08)		.003
Accuracy	5.13	.01*	.31 (SD= .17)	.42 (SD= .16)	.42 (SD= .16)	J vs. H F	.26
Complexity	17.77	.001*	1.24 (SD= .17)	1.43 (SD= .16)	1.28 (SD= .16)	H vs. J F	.55
Prop. time spoke	4.51	.05*	.83 (SD= .8)	.86 (SD= .7)	.87 (SD= .7)	J vs. F H	.24
Length of run	4.75	.05*	4.22 (SD= 1.91)	4.19 (SD= 2.37)	4.84 (SD= 3.16)	F vs. H	.25
Speech rate	.86	.43	101.54 (SD= 22.21)	102.51 (SD= 24.41)	104.81 (SD= 24.56)		.05
No. of pause mid	7.76	.006*	13.76 (SD= 8.22)	11.90 (SD= 6.79)	9.06 (SD= 5.50)	F vs. J	.35
No. of pause end	2.11	.14	4.86 (SD= 2.81)	5.40 (SD= 2.28)	4.30 (SD= 3.08)		.07
Total silence mid	2.82	.07	11.31 (SD= 9.13)	9.68 (SD= 7.20)	7.86 (SD= 5.94)		.16
Total silence end	3.37	.05*	3.80 (SD= 2.87)	3.54 (SD= 2.55)	2.58 (SD= 1.94)	F vs. J	.19
Pause length mid	.58	.54	.75 (SD= .22)	.75 (SD= .25)	.82 (SD= .31)		.04
Pause length end	1.90	.16	.75 (SD= .34)	.61 (SD= .22)	.62 (SD= .23)		.12

* Significant differences are reached.

Table: 9.23

Results of ANOVAs for Walkman, Picnic and Keys

Measures	*F*	*P*	Task / Structure: Walkman	Picnic	Keys	Sig. differences	Eta Squ.
Reformulation	2.64	.09	1.33 (SD= 1.72)	.83 (SD= .79)	1.16 (SD= 1.46)		.04
False start	9.20	.001*	6.23 (SD= 4.90)	3.63 (SD= 2.73)	5.36 (SD= 3.39)	P vs. W K	.39
Replacement	3.58	.09	2.86 (SD= 2.31)	1.90 (SD= 1.66)	2.60 (SD= 1.90)		.20
Repetition	.64	.53	5.46 (SD= 5.56)	4.63 (SD= 4.16)	4.93 (SD= 4.74)		.04
Accuracy	6.53	.009*	.30 (SD= .16)	.41 (SD= .21)	.41 (SD= .14)	W vs. P K	.31
Complexity	6.10	.009*	1.36 (SD= .27)	1.59 (SD= .38)	1.54 (SD= .24)	W vs. P K	.30
Prop. time spoke	.08	.91	.86 (SD= .7)	.86 (SD= .1)	.86 (SD= .7)		.006
Length of run	.60	.55	4.36 (SD= 1.71)	4.58 (SD= 2.33)	4.24 (SD= 1.54)		.04
Speech rate	.50	.60	96.84 (SD= 23.17)	99.06 (SD= 25.68)	95.99 (SD= 22.91)		.03
No. of pause mid	8.18	.001*	18.93 (SD= 13.54)	12.50 (SD= 9.86)	13.33 (SD= 7.23)	W vs. P K	.22
No. of pause end	7.47	.006*	8.80 (SD= 4.95)	7.16 (SD= 4.85)	6.03 (SD= 3.38)	W vs. P K	.20
Total silence mid	1.51	.23	16.79 (SD= 14.23)	13.15 (SD= 14.74)	12.64 (SD= 8.94)		.09
Total silence end	4.43	.05*	6.59 (SD= 4.72)	5.68 (SD= 5.19)	4.46 (SD= 3.70)	W vs. K	.24
Pause length mid	1.39	.26	.81 (SD= .27)	.98 (SD= .95)	.94 (SD= .94)		.09
Pause length end	.075	.92	.73 (SD= .24)	.70 (SD= .23)	.71 (SD= .33)		.005

* Significant differences are reached.

F = Football, H = Hunting, J = Journey, K = Keys, P = Picnic, W = Walkman

9.3 Results and the Hypotheses

9.3.1 Hypothesis 1

Hypothesis 1 predicted that language performance in tasks which present both foreground and background information will be more complex than performance in tasks which only present foreground information. T-tests are employed to explore whether any statistical differences exist between the performance in the foreground tasks as compared with performance in foreground and background information tasks. All the foreground and background information tasks, as revealed by the results of the t-tests, have elicited statistically more complex language than the foreground information tasks. With the unstructured tasks, Walkman with a mean ratio of subordination of 1.36 is significantly more complex than Journey with a mean of 1.24 (t = 2.01, P = .04*). For the structured schematic sequential tasks, Picnic with a mean ratio of subordination of 1.59 is significantly more complex than Hunting with a mean of 1.42 (t = 2.16, P = .03*). Similarly, for the structured problem-solution tasks, Keys with a mean ratio of subordination of 1.54 is significantly more complex than Football with a mean of 1.28 (t = 4.77, P = .001*). These results categorically support the prediction of the first hypothesis of the current study, indicating that performance in tasks which are based on both foreground and background information includes more syntactically complex language. As mentioned in Section 8.4.2, *vocd* measure of lexical variety was employed as an indicator of complexity of language performance. However, this measure did not show any systematic differences resulted from the grounding or task structure. As these results were difficult to be interpreted and appeared to refer to a different construct rather than complexity, they were excluded from further analyses and discussions.

As discussed before, fluency of performance was measured through a wide range of variables in the current study. False start, reformulation, replacement and repetition were the four measures of repair fluency. Numbers of pauses, length of pauses and total amount of silence were temporal aspects of fluency measured for both pauses happening within a clause and at clause boundaries. Speech rate, length of run and proportion of time spoken were other temporal fluency measures adopted to capture the detailed differences in performance across the tasks. In order to investigate whether grounding influences fluency of the performance, a number of t-tests were conducted on measures of fluency. Results of the t-tests show that different fluency measures have been influenced by the grounding variable. For instance, performance in foreground information tasks is significantly more fluent for measures of repetition and total amount of end-clause silence across all the tasks. Measures of false start, replacement and number of end-clause pauses are also significantly different for the unstructured tasks as well as the structured problem-solution tasks. This indicates that performance in tasks which present both foreground and background information, i.e. Walkman and Keys, has significantly fewer false starts and replacements and less silence than the performance in foreground information tasks, i.e. Journey and Football. For the number of mid-clause pauses, speech rate, total mid-clause silence, and length of mid-clause pauses the differences do not reach significance. However, in all these instances the foreground tasks are more fluent than the tasks with foreground and background information.

Independent-sample t-tests were further conducted to find out whether the accuracy of the participants' performance was influenced by the grounding variable. The results of the t-tests on the accuracy measure show that grounding does not influence accuracy of performance in tasks, since a significant difference is not reached between

any pairs of the tasks. This finding is in line with the previous predictions of the study: i.e. accuracy is influenced by task structure and not by grounding. In the following section, predictions of Hypotheses 2 and 3 of the current study, i.e. the effects of task structure, on various aspects of language performance will be investigated.

9.3.2 Hypothesis 2

This hypothesis predicted that language performance in the structured tasks would be more accurate than the performance in the unstructured tasks. This hypothesis was formed to confirm the findings of Study One and other previous research studies (Skehan and Foster, 1997). As the results of the ANOVAs show, for the foreground information tasks, a significant difference is observed among the accuracy measure of the three tasks ($F = 5.13$, $P = .01$*, $\eta^2 = .26$). The means of the accuracy measures for the three tasks clearly show that for the unstructured task, i.e. Journey, it is statistically less accurate than for both structured tasks (means of accuracy for Journey = .31, Hunting = .42, Football = .41).

Similarly, for tasks which present both foreground and background information, a significant difference is seen among the accuracy measure of the three tasks ($F = 6.53$, $P = .009$*, $\eta^2 = .31$). The means of the accuracy measure for the three tasks further demonstrate that the unstructured task, Walkman, is statistically less accurate than and thus different from the two structured tasks (means of accuracy for Walkman = .30, Picnic = .41, Keys = .41). It is worth mentioning that in the measure of effect size for both significant differences in the accuracy measure are noticeable, suggesting that a large amount of the variance emerged in the accuracy of the performances is associated with levels of the independent variable, i.e. task structure.

9.3.3 Hypothesis 3

As mentioned in the previous chapter, hypothesis three predicted that performance in structured tasks would be more fluent than performance in unstructured tasks. In both studies, fluency has been measured through a wide range of measures of repair fluency and temporal fluency including breakdown and speed fluency. The results of the ANOVAs on different measures of fluency support this hypothesis.

For measures of breakdown fluency, i.e. pauses and silence in the performance, in the foreground information tasks significant differences were observed for the number of mid-clause pauses ($F = 7.76$, $P = .002$*, $\eta^2 = .35$) and total amount of end-clause silence ($F = 3.37$, $P = .05$*, $\eta^2 = .19$). For both measures, performance in Football is significantly more fluent than in the other two tasks (means of number of mid-clause pauses for Journey =13.76, Hunting = 11.90, Football = 9.06; and means of total amount of end-clause silence for Journey = 3.80, Hunting = 3.54, Football = 2.58). This clearly shows that performance in structured tasks is progressively more fluent with less pauses and silence across the tasks. Interestingly, the estimate of the effect size for number of mid-clause pauses is noticeable.

With regard to length of run, a significant difference is further seen ($F = 4.75$, $P = .02$*, $\eta^2 = .25$) with performance in Football being significantly more fluent than and significantly different from performance in the other two tasks (means of length of run for Journey = 4.22, Hunting = 4.19, Football = 4.84). Comparison of the proportion of time spoken across the tasks indicates another significant difference ($F = 4.51$, $P = .02$*, $\eta^2 = .24$) with performance in the unstructured task being the least fluent and significantly different from performance in the structured tasks (means of proportion of time spoken for Journey = .83, Hunting = .86, Football = .87).

Regarding repair fluency measures, in the foreground information tasks, a significant difference is seen for measures of false start (F = 5.20, P = .01*, η^2 = .27) and replacement (F = 4.83, P = .01*, η^2 = .23). For false start, performance in one of the structured tasks, Football, is significantly more fluent than the performance in the other two tasks (Journey = 3.56, Hunting = 4.13, Football = 2.66). However, concerning replacement measure, performance in Journey is the most fluent (Journey = 1.13, Hunting = 2.23, Football = 1.26). For other measures of fluency in the foreground group a significant difference is not reached. However, for measures of reformulation, speech rate, total silence mid clause, and pause length end clause a linear relationship exists among the tasks with performance in the structured tasks being more fluent than in the unstructured task. This clearly indicates that performance in structured task is generally more fluent than performance in unstructured tasks. Hence, Hypothesis 3 receives general support from foreground information tasks.

Regarding the tasks that include both foreground and background information, results of the ANOVAs show that performance in the structured tasks is more fluent than in the unstructured task. For measures of breakdown fluency, significant differences are reached for number of mid-clause pauses (F = 8.18, P = .001*, η^2 = .22), number of end-clause pauses (F = 7.47, P = .006*, η^2 = .20) and total amount of end-clause silence (F = 4.43, P = .05*, η^2 = .24). Comparisons of the means of different tasks show that for number of mid-clause pauses, performance in the Walkman task is significantly less fluent than in Picnic and Keys (means for number of mid-clause pauses for Walkman = 18.93, Picnic = 12.50, Keys = 13.33). The same comparison for number of end-clause pauses reveal that performance in the unstructured task, i.e. Walkman, was significantly less fluent than and different from in the other two tasks

(means of number of end-clause pauses for Walkman = 8.80, Picnic = 7.16, Keys = 6.03). The same results are obtained for the total amount of end-clause silence with the Walkman task being significantly different from the other two tasks (means of end-clause silence for Walkman = 6.59, Picnic = 5.68, Keys = 4.46).

Regarding the false start measure of repair fluency, a significant difference ($F = 9.20$, $P = .001$*, $\eta^2 = .39$) is seen across the tasks with performance in Picnic being more fluent than in the other two tasks (means of false start for Walkman = 6.23, Picnic = 1.90, Keys = 2.60). For other measures of fluency a significant difference is not reached. However, in the case of repetition and total amount of mid-clause silence performance in the structured tasks is more fluent than in the unstructured task.

Surprisingly, the results of the ANOVAs show a significant difference on the measure of complexity across tasks between foreground tasks and foreground and background tasks. Based on the results of the statistical analyses on the complexity measure obtained from Study One, Study Two hypothesized that complexity would mainly be influenced by grounding and only marginally by task structure. However, the results of ANOVAs show a significant difference across the foreground information tasks ($F = 17.77$, $P = .001$*, $\eta^2 = .55$) with Hunting eliciting the most complex performance (means of complexity for Journey = 1.24, Hunting = 1.43, Football = 1.28). A further significant difference is observed for the complexity measure across the foreground and background information tasks ($F = 6.10$, $P = .009$*, $\eta^2 = .30$) with Picnic eliciting the most complex performance (means of complexity for Walkman = 1.36, Picnic = 1.59, Keys = 1.54). Interestingly, in both cases the unstructured task has exhibited the least syntactical complexity. Although these results do not directly reject any hypothesis of Study Two, they shed new light on the findings of the current research regarding the concept of complexity of performance.

A detailed discussion regarding the hypotheses and results of Study Two, an overview of the results and findings of Study One as well as the relevant discussions which relate to significant issues in SLA and LT literature will be presented in Chapter X.

CHAPTER X

Observations: Findings of Study One and Study Two

10.1 Overview

The overarching purpose of the current research was to explore the effects of task characteristics and performance conditions on the second language performance of Iranian language learners in an assessment setting. It attempted to find out whether there are cognitively demanding characteristics and performance conditions of tasks that influence task difficulty and consequently second language performance in oral narrative tasks. The first study investigated the effects of task structure, pre-task planning and language proficiency on different aspects of the language performance of 80 Iranian language learners of English. Study One further examined the effect of task characteristics and conditions on test-takers' perceptions of task difficulty. Based on the findings of the first study, a second study was developed to explore the effects of grounding on the complexity of the language performance of sixty Iranian language learners of English. Meanwhile, task structure was employed as an independent variable so that its effects on performance and its interactional effect with grounding on tasks could be examined. The principal objective of Study Two, therefore, was to investigate whether the cognitive demands of a task manifested in terms of grounding and task structure would influence the language performance of the participants on oral narrative tasks in terms of accuracy, complexity and fluency.

In the present chapter, discussions of the effects of grounding and task structure on performance will be first presented. Then a summary of the findings of Study One will be mentioned. The ensuing discussions will deal with the relationship between different aspects of performance in the current context of SLA. The discussion will then lead to an exploration of how the findings of this research could provide a wider perspective towards the interrelationship between grounding and task structure as well as how they influence language performance. In the last section of the current chapter, measures of fluency and complexity will carefully be discussed with regard to the relevant theories of language production.

To summarize, the results of Study Two clearly show that grounding influences complexity and fluency of performance in narrative tasks. The results clearly demonstrate that narrative tasks which contain both foreground and background information will elicit greater language complexity than the tasks which contain only foreground information. In addition to complexity, the fluency of performance is also influenced by grounding. In effect, as different fluency measures indicate, performance in tasks that present both foreground and background information is significantly less fluent than performance in foreground information tasks. With regard to the effects of task structure on the fluency and accuracy of language performance, the results of the current study confirm the findings of Study One, indicating that language performance elicited by structured tasks is more accurate and more fluent than the performance elicited by unstructured tasks. Task structure could also have a significant effect on complexity of performance. A prime finding of the present study concerns the intricate interrelationship between the three aspects of performance, i.e. accuracy, complexity and fluency, as different task characteristics

and the interaction among task characteristics influence them in slightly different ways. This interrelationship will be discussed later in this chapter.

10.2 Discussing the Findings of Study Two

10.2.1 Effects of Grounding

The effects of the grounding characteristic of tasks were investigated with respect to complexity, accuracy and fluency of performance. Based on the results of Study One, it was hypothesized that grounding would essentially affect complexity of performance. However, as a tradeoff relationship was expected between different aspects of performance, the effects of grounding were explored in a wider context for different measures of fluency, accuracy and complexity. Discussions in this section will focus on such effects on the three aspects of performance.

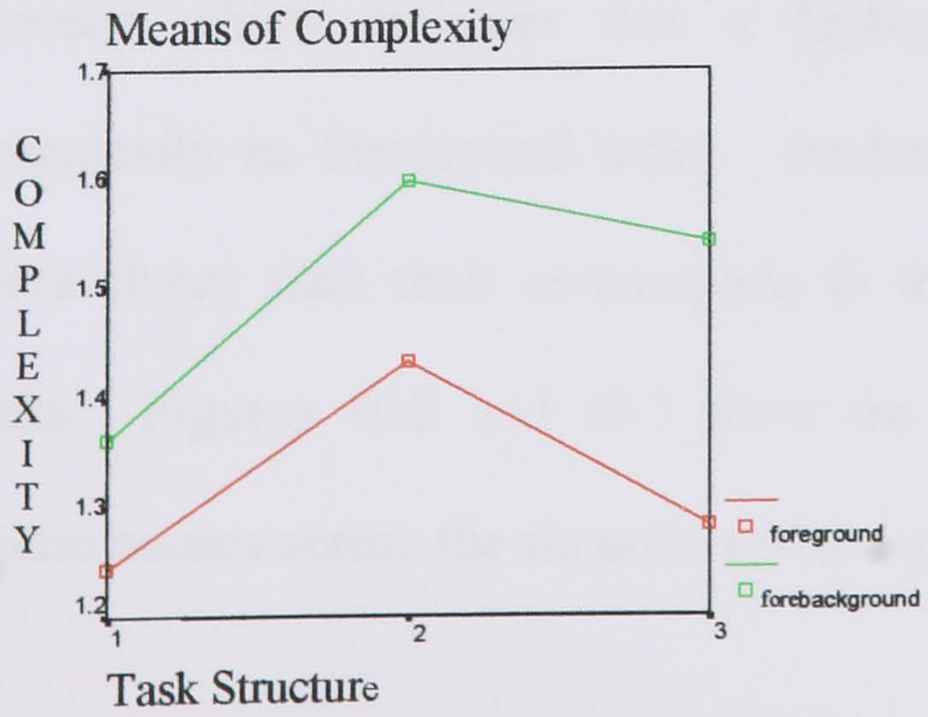
10.2.1.1 Complexity. In line with the theoretical motivations given earlier in chapter VIII regarding the effects of grounding on complexity, it was hypothesized that performance in tasks that contain both foreground and background information would be more complex than performance in tasks which contain only foreground information. This hypothesis is largely confirmed as the results of the t-tests show that all the foreground and background information tasks are significantly more complex than the foreground information tasks. Complexity, as defined by Ortega (2003), refers to the range of forms that surface in language production and the degree of sophistication of such forms. Skehan and Foster (1997) define complexity as a second language speaker willingness to attempt ambitious forms or to take risks by attempting less controlled language. It is worth mentioning that complexity, in this study, is measured through the ratio of subordination (subordinate clauses/ AS-units)

and therefore more complex language refers to language that includes higher ratios of subordination.

Results of the t-tests (Table 9.21) clearly show that presence of background information in a picture story stimulates the speakers to employ more subordination in their performance to fulfill the functional requirements of the task. It is evident that in performing a task which contains foreground and background information, the speaker has a tendency to be ambitious and to use more complex language to show the events occurring in the foreground, relate them to the stories happening in the background and describe the relationship between the two. It may be interpreted that the simultaneity of the events in the foreground and background settings motivates the speaker to employ a number of subordinating clauses including cause and effect and time clauses. It could be argued that, when the background information is eloquently adjoined to the foreground information, it creates a rich context for the narrative. The speaker will then need more complex language to weave the background information into the foreground events to demonstrate the rich context of the narrative thoroughly. As a reminder, Figure 10.1 demonstrates the means for complexity of performance across the six tasks, their task structure and grounding.

Figure 10.1

Complexity: Effects of Grounding



1= Unstructured 2= Schematic Structured 3= Problem-solution Structured

10.2.1.2 Fluency. Meanwhile, comparisons of the fluency measures show that grounding greatly influences different fluency measures as well. The results clearly indicate that performance in foreground tasks is more fluent than performance in foreground and background tasks with statistically less silence and fewer repetitions for all the foreground tasks. In effect, the results explicitly indicate that when the performance in a task is more complex less fluency is generally associated with this high complexity. This clearly supports the findings of previous research, particularly in cognitively-oriented task-based research (Skehan, 1998), claiming that attentional resources are limited and that to attend to one aspect of form may mean that other dimensions would suffer. This view to language processing is mainly proposed by cognitive psychology and stresses that attention is both limited and selective (McLaughlin, Rossman & McLeod, 1983). This suggests that, as participants put more effort in achieving greater syntactic complexity in their performance, or as they pay more attention to the form, they have fewer attentional resources available to adhere to the fluency of their performance.

With regard to other measures of fluency, the same competing relationship between complexity and fluency is pervasive. The results show that grounding does not significantly influence other measures of fluency, i.e. length of run, speech rate or pause length. However, there is a clear and noticeable trend across all measures of fluency which indicates that a higher level of fluency is associated with less complexity in foreground tasks. Predominantly, Journey, Hunting and Football are more fluent than their counterparts in the foreground and background information tasks. Figures 10.2 and 10.3 show the means for speech rate and number of end-clause pauses across the six tasks and the grounding groups.

It should be noted that the scores for different fluency and complexity measures of Hunting are different from the scores obtained on Football and Journey. As discussed in Chapter VIII, Hunting was not a purely foreground task since some implicit background information was represented in the pictures. As a result, it was anticipated that performance in Hunting would be slightly different from performance in the other two foreground information tasks. This prediction was confirmed when more complexity was generated in the performance in Hunting than on Journey and Football. The low amount of fluency in Hunting, therefore, appears to be a consequence of producing more complex language as required by the background information presented in Hunting. This, in turn, supports the effects of grounding on different aspects of performance and indicates how these different aspects interact in a tradeoff relationship.

Figure 10.2

Speech Rate: Effects of Grounding

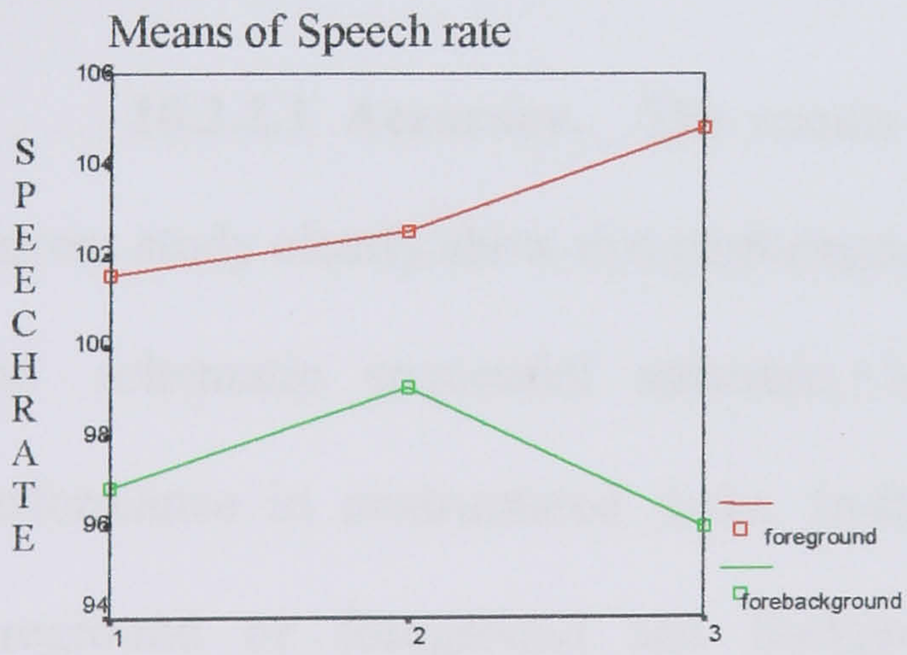
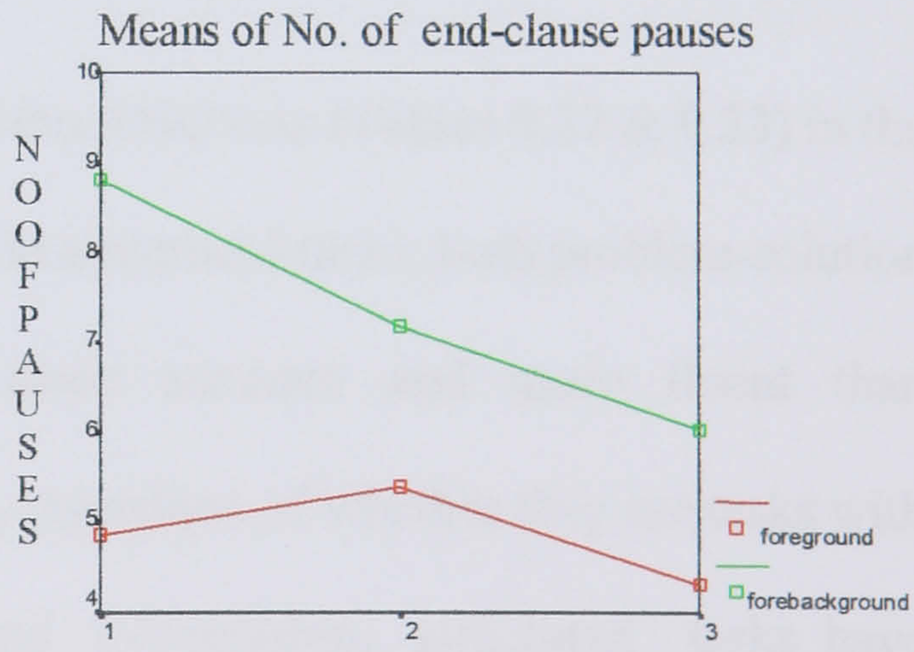


Figure 10.3

Number of End-Clause Pauses: Effects of Grounding



10.2.1.3 Accuracy. As regards the accuracy measure, it was hypothesized that grounding would not influence accuracy. The results of the t-tests (Table 9.21) on the accuracy measure confirm this prediction as there is no significant difference between pairs of tasks which are different in the type of their grounding. In other words, tasks

which contain foreground information, are not significantly different from tasks which contain both foreground and background information. The discussions presented in this section attempt to explore the findings of this research regarding the effects of grounding. However, as these results interact with the results obtained from the other independent variable of the study, i.e. task structure, the interrelationship between the two independent variables and how they would have an overall effect on performance will be discussed later in this chapter.

10.2.2 Effects of Task Structure

In line with the findings of previous research as well as the results of Study One, it was hypothesized in Study Two that the inherent structure of a task would enhance the accuracy and fluency of performance. However, a direct effect on complexity of performance was not anticipated. In the section that follows, the effects of task structure on accuracy, fluency and complexity will be discussed.

10.2.2.1 Accuracy. The results of the ANOVAs (Tables 9.22 & 9.23) in the current study clearly show that performance in structured tasks, both problem-solution and schematic sequential structure, is more accurate and more fluent than performance in unstructured tasks. In fact, regardless of whether they are tasks with foreground or foreground and background information, structured tasks have generated more accurate and more fluent language. Hence, this study replicates the findings of previous studies (Skehan & Foster, 1997, 2001; Wigglesworth, 2001) and supports the results obtained from Study One.

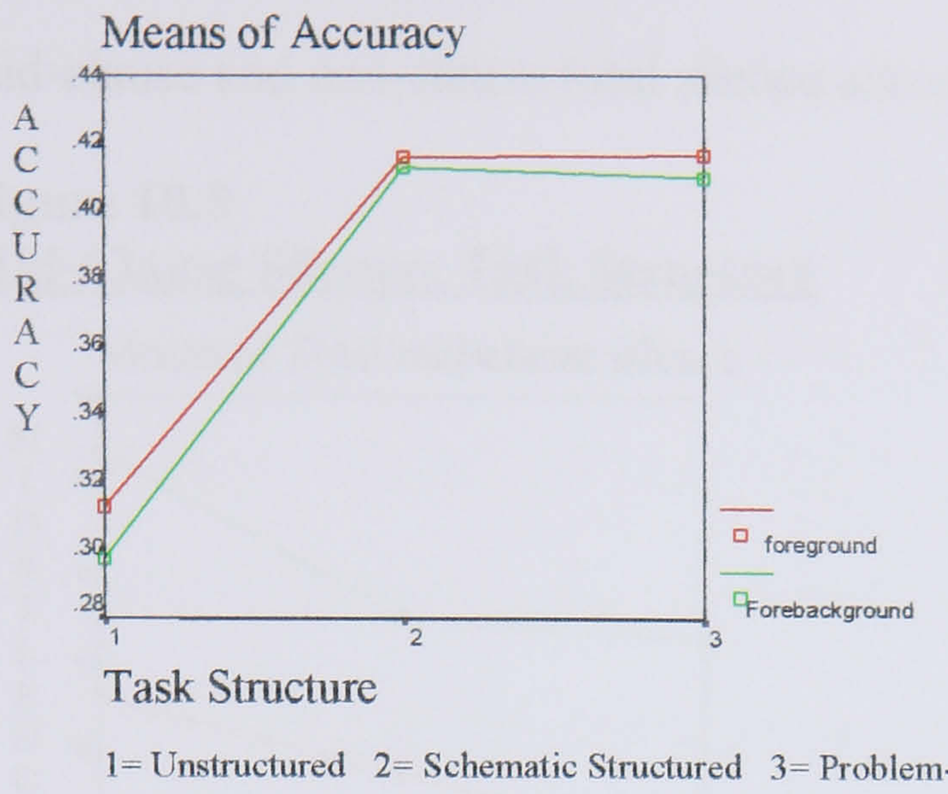
For both grounding groups of tasks, performance elicited by the unstructured task is statistically less accurate than performance elicited by both structured tasks. In

foreground tasks, Journey lacked structure as there was not a clear time line in the narrative and the events were loosely related. On the other hand, Hunting and Football were based on schematic and problem-solution structures respectively. Similarly, for the tasks which had both foreground and background information, Walkman was unstructured, whereas Picnic and Keys had schematic and problem-solution structures respectively. Interestingly, similar to the results of Study One, performance in the structured tasks has elicited equal means on the accuracy measure for both grounding groups of tasks. In effect, the unstructured tasks have elicited performance that is less accurate than in the structured tasks, whereas the structured tasks have all elicited more accurate performance with similar figures to one another for the accuracy measure obtained in Study One.

In line with the findings of cognitive psychology regarding the allocation of attentional resources, it could be argued that lack of structure in a task increases the cognitive demands of the task and consequently requires the participant to have more attentional resources available to perform the task. On the other hand, as participants are constrained with a limited amount of attentional resources, they are less likely to pay adequate attention to different aspects of their performance. In contrast, for structured tasks, the participant perceives the clear inherent structure of the task, which would make performing the task less demanding. Not being concerned with lack of structure, therefore, the participant has the opportunity to attend to different aspects of form, i.e. accuracy and complexity, as well as meaning, i.e. fluency, while performing a structured task. Figure 10.4 shows the means of accuracy for different tasks and grounding groups.

Figure 10.4

Accuracy: Effects of Task Structure



10.2.2.2 Fluency. With regard to fluency, performance in the structured tasks is also more fluent than the unstructured tasks on a variety of fluency measures. For instance, performances elicited by Football and/or Hunting are statistically more fluent than performance elicited by Journey on measures of proportion of time spoken, number of mid-clause pauses and end-clause pause length. In tasks with foreground and background information, performances elicited by Picnic and Keys are statistically more fluent than performance elicited by Walkman on measures of false start, number of mid-clause pauses, number of end-clause pauses and end-clause total silence. For many other measures of fluency, although statistical significance is not reached, it is clearly seen that performance in the structured tasks is more fluent than performance in the unstructured tasks on different fluency measures. This further reveals how the cognitive demands of a task influence the performance in tasks, not only in terms of accuracy but also regarding fluency.

These results, in effect, clearly suggest that the presence of structure in a task makes the task look less demanding to the participants who would therefore be able to employ all their attentional resources to produce more fluent language. In contrast, with the unstructured tasks, it appears that the participants' attentional resources are

partly employed to make up for the lack of structure and participants would only partly attend to fluency of their performance. Figure 10.5 and 10.6 show the means of mid-clause and end-clause total silence across different tasks and grounding groups.

Figure 10.5
Mid-Clause Silence: Task Structure

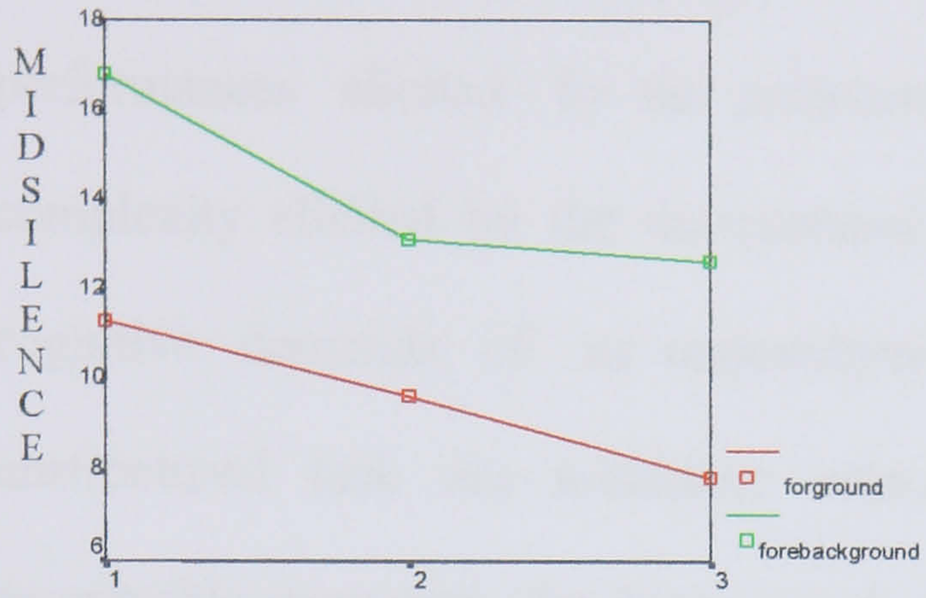
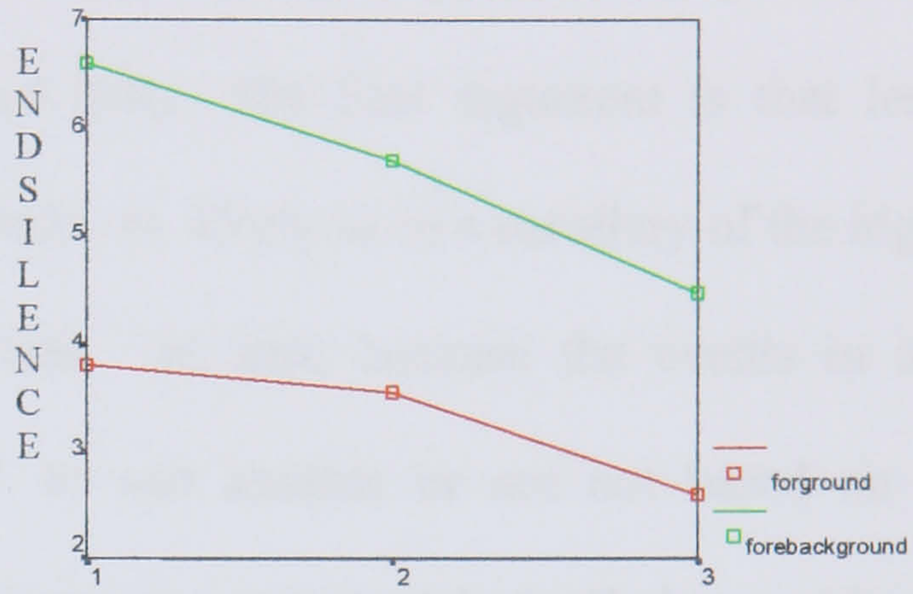


Figure 10.6
End-Clause Silence: Task Structure



1= Unstructured 2= Schematic Structured 3= Problem-solution Structured

10.2.2.3 Complexity. Regarding complexity, a direct effect of task structure on language performance was not predicted. However, the results generally show that the structured tasks are more complex than the unstructured tasks (Figure 10.1). Structured tasks, in fact, have either at a significant or at a non-significant level elicited greater complexity than the unstructured tasks. In the case of foreground information tasks, Hunting and Football are more complex than Journey, and among tasks with foreground and background information Keys and Picnic are statistically more complex than Walkman.

It was discussed earlier that Hunting, because of its embedded implicit background information, was expected to elicit more complex language than the other two foreground information tasks. In fact, as explained in Chapter VIII, while examining different picture stories for Study Two Hunting was recognized as a foreground information picture story with an implicit background that could influence performance. However, because of the practical restrictions of the study Hunting was

employed to serve as a foreground information task. Hence, the high complexity of performance elicited by hunting can be connected to this implicit background information provided in the pictures. However, the high complexity of performance in Picnic and Keys was only partly anticipated.

Two arguments could be represented to explain the low syntactic complexity of performances elicited by the unstructured tasks. The first argument is that less complexity elicited by the unstructured tasks is likely to be a corollary of the high cognitive demands of an unstructured task. In fact, because the events in an unstructured task are arbitrarily related to one another or are not based on a perceivable sequence, the background information cannot realistically be combined with the foreground events. In other words, certain events occur in the background, whose relevance to the foreground events tends to look trivial to the participants. As the foreground events are themselves loosely related to one another, it seems more difficult to the participant to consider all the occurring events and relate them together in an integrative way. Alternatively, it could be argued that there are other latent task characteristics intrinsic to some tasks, i.e. Picnic and Keys, which are not taken into account in the present research. Obviously more research is required to provide more evidence on why some structured tasks have elicited greater language complexity.

10.3 A summary of the Findings of Study One

Study One attempted to investigate whether degree of task structure, pre-task planning conditions and language proficiency have significant influences on different aspects of the performance of Iranian test-takers of English and on their perceptions of task difficulty. Following the literature from which the theoretical motivations of the study were drawn, four oral narrative tasks were employed in the study. Task

structure was, for the first time in task-based research, thoroughly investigated and subsequently defined in terms of two types of structure frequently mentioned in SLA literature, i.e. problem-solution and schematic sequential structure. The problem-solution structure was assumed to be a stronger structure type and to have a greater effect on facilitating performance than the schematic sequential structure. Both types of structure were assumed to be, in turn, more structured than the other two tasks, i.e. Walkman and Unlucky man which were arbitrary sequences of events. Planning time was operationalized through providing either 5 minutes or 30 seconds of pre-task planning time to the participants before performing the tasks. Participants were selected from two levels of language proficiency, i.e. elementary and intermediate so that any interaction between language proficiency, task characteristics and conditions could be explored. Finally, to understand how participants of the study perceived task difficulty, perceptions of the participants on the difficulty of the four tasks were investigated by use of retrospective questionnaires.

The results of Study One clearly indicate that task structure has a positive influence on accuracy and fluency of performance. Performance elicited by structured tasks proves to be, in line with the predictions of the study, progressively more accurate and fluent on the structured tasks. However, task structure does not appear to have a systematic influence on syntactic complexity of the language performance. Pre-task planning conditions positively influence language performance in all aspects of accuracy, complexity and fluency. Planned performances, in effect, have proved to be more accurate, more complex and more fluent than unplanned performances for both elementary and intermediate language proficiency groups. The improvement of fluency that emerged as a result of providing planning time to the participants was even greater than the effect of language proficiency on some fluency measures.

Investigations of the participant's perceptions of task difficulty clearly demonstrated that participants perceived the unstructured tasks as significantly more difficult than the structured tasks under both planning conditions. A summary of the effects of task structure on different aspects of language performance is provided in Table 10.1.

Table 10.1

A Summary of Effects of the Task Structure in Study One

Structured Tasks		**Unstructured Tasks**	
Football	*Picnic*	*Unlucky Man*	*Walkman*
More accuracy More fluency Less complexity Perceived as less difficult	More accuracy More fluency More complexity Perceived as less difficult	Less accuracy Less fluency Less complexity Perceived as more difficult	Less accuracy Less fluency More complexity Perceived as more difficult

These results are obtained in Study One where task structure is the only task characteristics. However, in Study Two, two task characteristics have been considered and manipulated in each of the six picture stories, i.e. grounding and task structure. In the following section, therefore, the discussions will deal with the relationship between the three aspects of language performance, i.e. accuracy, complexity and fluency, when grounding and task structure combine with each other and play certain facilitative or restrictive roles in the participants' performances in the tasks.

10.4 Discussing the Overall Findings of the Two Studies

10.4.1 The Relationship between Different Aspects of Language Performance

A recent controversy in the field of second language acquisition in general, and task-based research in particular, expresses two contrasting approaches regarding the interrelationship between the three aspects of performance, i.e. accuracy, complexity and fluency. It should be noted that in this context, accuracy and complexity of the

language performance are taken to represent the *form*, while fluency is taken to indicate the prioritizing of meaning by language learners, i.e. getting communication across without worrying about its form (Foster & Skehan, 1996; Skehan & Foster, 1997). Within the cognitive approach to task-based instruction, the primary argument concerns whether attending to form, i.e. accuracy and complexity, would mean that there would be less attention to meaning, i.e. fluency. In other words, it is not clear yet whether there is a tradeoff relationship between form and meaning. It also raises the question of whether one aspect of form might compete with the other aspect and therefore result in a deterioration of the performance. Skehan (1998, 2001) suggests that accuracy and complexity are in competition for attentional resources. He, following VanPatten (1990), argues for a limited attentional capacity – that more difficult tasks require more attention to the content of the task and this will subsequently impose limitations on availability of attention to form. It has been argued that, because there is a limited supply of attention and because an activity that draws upon attentional resources will interfere with other activities requiring it, attention must be strategically allocated (Shaw & Shaw, 1978).

In contrast, Robinson (2000, 2001) argues that form and meaning need not always be in competition for scarce attentional resources. He reports from other researchers (Navon, 1989; Neumann, 1996) who contend that limited capacity and single-resource models of attention are questioned by recent research into task performance. Robinson, in effect, claims that different aspects of performance do not need to compete with one another. In other words, he contends that, difficult tasks would elicit language performance which is fluent, accurate and complex. In this section, the discussion will focus on the results obtained from the current research and how they relate to this controversy in SLA research.

The results of the factor analyses in Study One clearly indicate that accuracy and complexity have a high association across the tasks. For all the four tasks, accuracy and complexity loaded on the same factor and contrasted with measures of fluency which loaded on the other factors. This provides evidence that, in all instances, fluency competes with the two aspects of form, i.e. accuracy and complexity. Furthermore, the negative correlations between different measures of fluency on the one hand and accuracy and complexity on the other hand confirm the existence of a tradeoff relationship between form, i.e. accuracy and complexity, and meaning, i.e. fluency. In fact, for all four tasks accuracy and complexity have large negative correlations with silence and repair measures of fluency. For three of the tasks - Football, Picnic and Walkman, accuracy and complexity show large positive correlations, revealing no competition between the two. However, for Unlucky Man the correlation between accuracy and complexity is marginal, suggesting that a different type of relationship may exist between accuracy and complexity (See correlation matrices in Chapter VI).

Study Two is a further development of Study One, in which more independent variables are employed to explore the relationship between task characteristics and their effects on aspects of language performance in general and on complexity in particular. The main hypotheses of Study Two predicted that grounding would influence complexity while task structure would have effects on accuracy and fluency of performance. The results of the different analyses, including the factor analyses, in Study Two have demonstrated that grounding clearly influences different aspects of performance.

As a general finding of Study Two, it is clear that performance in foreground tasks is more fluent and less complex. In fact, as the tasks do not produce high levels of

syntactic complexity, more attention is available to be paid to fluency. Therefore, there is an initial tradeoff between complexity and fluency across all the tasks. For foreground information tasks, the results of the factor analyses show that the two aspects of form, i.e. accuracy and complexity, constantly load on two distinct factors indicating a weak association between the two variables across the tasks. In effect, for Journey, Hunting and Football, accuracy and complexity bear small or slightly negative correlations (correlation coefficients for Journey: r=.150, p< .428, Hunting: r =.137, p< .471, Football r =-.06, p< .736). This refers to a lack of association between the two components of form and reflects another tradeoff relationship between accuracy and complexity in foreground information tasks. The results of the factor analyses and the correlations, in fact, suggest that priority is given either to accuracy or to complexity, and fluency is a corollary of how attention is divided into these two aspects of form (See Tables 9.7-9.12 for the correlation matrices). These results are, to some extent, different from the results of Study One. The difference could be attributed to either the smaller number of participants performing each of the tasks in Study Two or to the intricate interactions between task structure and grounding. More systematic research is required to investigate the relationship between accuracy and fluency of language performance in tasks in detail.

This finding supports Skehan and Foster's (1997, 2001) argument on the existence of tradeoffs in performance, as greater fluency may be accompanied by greater accuracy or greater complexity, but not both. On the other hand, this finding is in contrast with the previous arguments in SLA, claiming that limited attentional resources are directed first towards those elements that carry message meaning and only at a later stage towards redundant formal features of language (VanPatten, 1990, 1994; Lee, Cardino, Glass, & VanPatten, 1997). In effect, the later discussions in this chapter

will show how it is likely that certain aspects of form receive primary attention under certain circumstances.

The results of the factor analyses for tasks that contain both foreground and background information introduce a different picture about the interrelationship between the three aspects of performance. Unlike the results from the foreground information tasks, complexity and accuracy of performances on foreground and background information tasks constantly load on the same factor across the three tasks. This association of the two variables proposes that with foreground and background information tasks, the higher levels of complexity are incorporated with more accuracy in the participant's performance. Moreover, positive correlations are observed between accuracy and complexity measures across the three tasks with two of them having significant values (Pearson correlations for Walkman: r =.371*, p < .03*; Picnic: r =.356, p < .06; Keys: r = .521*, p < .003*). The figures indicate that for these tasks complexity and accuracy do go together to a larger extent. It seems that the presence of background information along with the foreground events pushes the participants to utilize more complex language, while they are paying equal attention to accuracy as well. In other words, there seems to be an integrative function in the background information which helps participants generate complex and accurate language simultaneously.

This finding confirms Robinson's (2001) argument that language learners have access to multiple and non-competing attentional pools. Robinson, following Givon (1985), contends that complexity and accuracy of performance in tasks correlate, as they are driven by the nature of the functional linguistic demands of the tasks. The findings of Study Two regarding tasks that contain both foreground and background information confirms Robinson's claim, while the results of the factor analyses on the foreground

information tasks confirm Skehan's claim about the tradeoff relationship between accuracy and complexity.

The evidence provided in this section directly touches upon the complicated nature of the relationship between fluency, accuracy and complexity. The results from both studies strongly confirm a tradeoff relationship between form and content. Moreover, the results provide clear evidence on the existence of another tradeoff between the two aspects of form under certain circumstances. With foreground information tasks accuracy and complexity compete with one another. However, when tasks are cognitively demanding because the background information is being related to the foreground information, accuracy and complexity go hand in hand. It appears that grounding plays a significant role in allocating the attentional resources and in influencing language performance.

In the previous section, the effects of grounding on language performance were discussed. However, when task characteristics interact with one another a combined set of effects on language performance could be expected. The discussions of how different task characteristics interact on one another and how they separately and jointly influence the three aspects of performance will be presented in the next section.

10.4.2 The Interrelationship between the Effects of Grounding and Task Structure

The overall results of Study Two, regarding the interrelationship between the two independent variables, how they interact with one another and thus influence different aspects of performance is summarized here. It is clearly evident now that task structure and grounding significantly influence language performance. Presence of

structure in a task appears to elicit more accurate and more fluent language, whereas type of grounding in a task influences complexity and fluency. However, it should be noted that these two characteristics do not exist separately in tasks. Nor do they act in isolation. They appear to interact with each other while the participants are engaged in the process of production and, for this reason, language performance tends to be influenced by both. It is also essential to consider this interaction inasmuch as the cognitive demands of tasks and the participants' limited attention are concerned.

The results have indicated that foreground information tasks do not elicit highly complex language since there is not background information which needs to be incorporated into the main events of the story. As complex syntactic structures are not required by these tasks the participants do not have to pay much attention to this aspect of form. Hence, the tasks will not be demanding as far as attentional resources are required. Therefore, the participant will be able to devote more attention to fluency. On the other hand, as the results of both studies show, accuracy is predominantly influenced by structure. Presence of a clear structure, whether problem-solution or schematic sequential, facilitates the production processes by freeing up the attentional resources for the participant to deal with accuracy and fluency. In contrast, with tasks which contain both foreground and background information attentional resources become scarce because the nature of the background information in the task, as discussed before, requires the speaker to produce greater language complexity. Ultimately, the participants' attempts to produce more complex language will reduce the amount of attention available to be paid to fluency. However, if the task is structured more accuracy and more fluency will be generated since the framework of structure has a positive influence on the performance of the speaker.

Taking the interactional effects of grounding and task structure into account, the performance of the participant on each individual task would require a different account. The Journey task elicits a low amount of complexity, as the foreground requirement of the task does not impel the participant to use higher levels of syntactic complexity. But the lack of background information in Journey, which implies that there is no necessity to use complex language, is likely to free up some attentional resources to the participant. Therefore, the participant's attention is not consumed and can be used more attentively in performing the task fluently. On the other hand, as Journey is an unstructured task, lack of structure increases the cognitive demands of the task and thus diminishes the resources available to deal with accuracy. As a result, performance in Journey becomes less complex, less accurate but relatively more fluent. In fact, performance in Journey becomes more fluent than the performance in tasks whose background information requires more complex language- Walkman, Picnic and Keys- but less fluent than performance in foreground tasks with a clear structure, i.e. Hunting and Football.

Hunting and Football are not cognitively demanding, since they have an inherent task structure, and do not elicit complex language, because they do not contain background information. Therefore, more attention can be paid to fluency. Since they have a clear task structure, both tasks would elicit performances that are more accurate as well. Hence, performances in Football and Hunting present low complexity but elicit high fluency and accuracy.

Walkman as a task with foreground and background information will primarily elicit greater syntactic complexity, which would in turn impose restrictions on the attentional resources available to the participant. Not having access to enough attentional resources, the participant will have to pay less attention to fluency. In

addition, as Walkman is an unstructured task, lack of structure increases the cognitive demands of the task. Subsequently, the participant will not be able to pay enough attention to accuracy and fluency. As a result, performance in Walkman is the least fluent because the cognitive demands of Walkman is high both in terms of the lack of structure and the requirements of foreground and background information.

Picnic and Keys are also tasks with foreground and background information and thus will elicit more complex and subsequently less fluent performance. However, as they are structured tasks, no further restriction is imposed on the attentional resources available to the participants performing these tasks. So performance is likely to be more accurate and more fluent. The interactional effects of different task characteristics would lead the performance in Picnic and Keys to be highly complex and accurate. Performance in Picnic and Keys is more fluent than performance in Walkman since Walkman suffers from the lack of structure. However, performance in Picnic and Keys is less fluent than performance in all the foreground tasks, which have free attentional resources available, as they do not require syntactically complex language. A summary of the interactional effects of different task characteristics on aspects of language performance in the current research is shown in Table 10.2.

Table 10.2

Effects of Task Characteristics on Language Performance

	Structured	**Unstructured**
Foreground Task	More accurate More fluent Less complex	Less accurate Less + More fluency Less complex
Foreground/background Task	More accurate More complex Less fluent	Less accurate More complex Less fluent

A significant feature of the current research has been the wide range of fluency measures adopted in both studies. The various measures of fluency employed in this research, how they function across different tasks and how they relate to other aspects of performance are issues to be discussed in the section that follows.

10.4.3 Fluency Measures

A principal characteristic of the current research is the attention it has paid to different measures of fluency. Practical restrictions of measuring fluency on the one hand and the multifaceted nature of fluency on the other hand have, in the past, prevented task-based research from exploring the complicated nature of fluency as a significant construct in language performance. The availability of computerized software and digital technology has recently provided researchers with an opportunity to tackle the intricate problems of measuring different aspects of fluency. Fluency was traditionally measured through the two categories of temporal and repair fluency. Skehan (2003) argues that analyses of fluency require separate measures of (a) breakdown fluency, or silence, (b) repair fluency, (c) speech rate, and (d) automatization of performance through length of run. He states that these sub-dimensions of fluency are needed if a comprehensive picture of performance is required.

Following Skehan (2003) and to meet the need of a careful investigation of fluency in the context of task-based research, it was decided that various aspects of fluency would be measured in the current research. In Study One, ten different measures of reformulation, false start, replacement, repetition, length of run, speech rate, total amount of silence, number of pauses, mean length of pauses and proportion of time spoken were employed. The results of the factor analyses from Study One revealed

that total amount of silence, mean length of pause, number of pauses, length of run, speaking time and speech rate loaded on the same factor. These measures, which can briefly be called temporal measures, correlate highly with each other and appear to refer to a single construct. Through all the different statistical analyses in Study One, the temporal measures have indicated a high degree of consistency and inter-relatedness across tasks and performance conditions. Although some of the temporal fluency measures did not reach significant differences as influenced by task structure, performance in structured tasks and under planned conditions was consistently more fluent for all temporal measures of fluency.

In contrast, the results of the analyses on the repair fluency measures in Study One seem to be very different. Initially, the four measures of false starts, reformulations, repetitions and replacement loaded on the same factor across the four tasks, suggesting that they represent the same underlying construct. But, by considering the correlation coefficients between the temporal and repair fluency measures, it can be noted that there are hardly any large correlations between the repair and the temporal measures of fluency. Nor is there a correlation among the repair measures themselves. In fact, the only high correlation between the repair fluency measures is between false starts and reformulations (r = .82). It is worth mentioning that this high correlation is explained by the fact that false starts themselves are a pre-requisite to reformulations, that is reformulations can only occur after false starts are made. No clear pattern of increase or decrease was observed, either as a function of task structure or planning time, in repair fluency measures. Such results could suggest that false starts are a normal part of any oral language performance, and therefore will not be directly influenced by planning. Alternatively, it could be argued that larger amount of silence provides the speaker with an opportunity to avoid false starts.

Predictably, false starts and reformulations have shown significant differences resulting from the effect of proficiency level. A trend towards increased fluency was only observed for replacement and repetition in the comparison between low to high-proficiency levels.

Reformulations are generally considered as part of the process of repairing performance, or on-line processing, in the current study. Participants abandon what they have said and try to reformulate it. However, it is not clear whether participants employ reformulations to correct, complete or intensify utterances.[1] Nor are the reformulations evaluated in terms of the achieved accuracy or success of the outcome of the performance in the data analysis. Since adequate evidence is not available from the results of the studies, it could be argued that there are cognitive processes involved in employing repair fluency, which are not manipulated within the restrictions of the present study. Finally, with the limitations of the present study it cannot be concluded whether repair fluency measures exclusively reflect patterns of fluency or are mainly a function of other unforeseen cognitive processes.

A principal development of Study Two has been recruiting more measures of fluency to have a more in-depth look at the aspects of fluency. As explained in chapter VIII, in Study Two the number of pauses is considered separately for pauses of longer than .4 of a second occurring in the middle of clauses as contrasted with those happening at clause boundaries. This distinction is primarily made to explore more about different aspects of fluency and at an advanced level to investigate whether mid-clause or end-clause pauses would have a greater effect on fluency. In so doing, it gave the researcher the opportunity to find out how mid-clause versus end-clause pauses relate

---

[1]Investigations of the data demonstrate that speakers sometimes utilize reformulation to make the performance more accurate and sometimes to make it more complex. They also employ reformulations to show a change in their decisions.

to other measures of fluency as well as to other aspects of performance. In this regard, a salient feature of Study Two is exploring the relationship between speech rate and mid-clause and end-clause pauses.

Speech rate, in this research, is assumed to define the speed with which a task is performed and is calculated by dividing the total number of syllables produced in the performance by the amount of time expressed in seconds. Obviously, the number of pauses occurring in a performance would negatively correlate with the speech rate of the performance. The general results of the factor analyses indicate that these three measures of fluency have loaded on one factor. Unsurprisingly, the Pearson correlation coefficients have shown negative correlations between speech rate and number of end-clause pauses (Journey: r= -.388, p< .03*, Hunting: r= -.351, p< .06, Football r = -.131, p< .49; Walkman: r = -.095, p < .61; Picnic: r = -.377, p < .04*; Keys: r = -.349*, p < .06*). Strikingly, however, for all the tasks a higher significant negative correlation is seen between the number of mid-clause pauses and speech rate (Journey: r = -.569*, p< .001, Hunting: r = -.611*, p< .001, Football r=-.541, p< .002*; Walkman: r = -.522*, p < .003*; Picnic: r = -.627*, p < .000; Keys: r = -.674*, p < .000*). These results firmly propose that pauses occurring in the middle of clauses have higher correlations with the speed of performance. In other words, it is reasonable to claim that what makes a performance less fluent depends, to a large extent, on the number of mid-clause pauses. These results indicate that the speed of performance is considerably influenced by pauses occurring in the middle of clauses.

As discussed in SLA literature (Oppenheim, 2000; Stern, 1992), nativelike delivery of English has certain features, one of which is short pauses of less than .4 seconds between short stretches of speech. In fact, different studies of native speakers' fluency have shown that native speakers of English normally pause between, and not

within, stretch of speech, which in the case of the current research refers to clauses. Therefore, the findings of this research suggest that the dysfluency in second language learner's speech, or what distinguishes a second language speech from that of a native speaker, is greatly related to the mid-clause pauses.

As explained in the previous chapter, to avoid violating theoretical assumptions of factor analysis, other measures of breakdown fluency were excluded from the analyses. However, correlations between speech rate and other measures of breakdown fluency, i.e. mid and end-clause silence and mid and end-clause pause length, show a similar pattern. For all these measures, the negative correlations between speech rate and mid-clause pauses are constantly higher than the correlations between speech rate and end-clause pauses (See Appendix 7 for the correlations among the different measures employed in Study Two). This further confirms that the main breakdown in second language performance results from the breakdown fluency in the middle of a clause.

10.4.4 Complexity Measure

Results of Study One demonstrated that complexity of the performances was not influenced by task structure. Based on this finding, Study Two was designed to investigate which task characteristics would lead the performance in a task to become more syntactically complex. The results of Study Two explicitly demonstrate that background information in an oral narrative task will influence performance in the task, particularly its syntactic complexity. To have a more theoretical discussion on the effect of task characteristics – grounding and degrees of task structure - on different aspects of language performance, particularly on complexity, Levelt's (1989, 1993) language production model is considered here.

Levelt (1989) proposes that the processes of speech production fall into three broad areas of conceptualization, formulation and execution. He proposes that the most fundamental level is the conceptualization stage and involves processes which determine what the speaker intends to say. The second stage is formulation which involves processes translating the conceptual representation into linguistic forms. Finally, the execution stage involves detailed phonetic and articulatory planning of the language to be produced. As regards the purpose of this research, conceptualization and formulation processes are more significant since they appear to influence accuracy, complexity and fluency of performance.

As Levelt contends, conceptualization deals with proposing and generating ideas which will be produced at a later stage. During conceptualization, speakers collect information and formulate ideas in the preparation of constructing what they intend to say. Levelt (1989) calls this stage the message level, which involves organizing and sequencing of the ideas. Distinguishing between macroplanning and microplanning processes within the conceptualization, Levelt (1989) argues that microplanning involves assigning the right form of language to different chunks of information in order to achieve the communicative purposes of the utterances.

Formulation, on the other hand, includes the two major processes of lexicalization and syntax. The propositional messages would move from the conceptualization to the formulation and would be subjected to grammatical and phonological encoding. Syntax would consequently emerge from the lexical elements required by the conceptualization processes. Therefore, it can be inferred that complexity of performance is formed as a result of the processes involved in conceptualization, whereas accuracy and fluency are developed as a result of the processes involved in formulation.

The results of the current research confirm the model but shed a new light on some aspects of Levelt's model. Complexity, as discussed within Levelt's (1989) model, is different from fluency and accuracy since it is constructed in a different stage and results from different processes. In Study One, it was hypothesized that task structure would influence fluency and accuracy. It was also assumed that existence of structure in a task would reduce the cognitive load and thus provide the participants with an opportunity to pay more attention to different aspects of form and meaning. In other words, it was postulated that conceptualization processes would benefit from inherent task structure and this would consequently facilitate the formulation processes which would result in an increase in accuracy and fluency.

The results of both studies showed that complexity is mainly influenced by task characteristics other than task structure. Although task structure influences language performance, it does not seem to be mainly activating the processes which would lead into producing more complex language. The results of Studies One and Two support Levelt's model, claiming that the first stage in production, i.e. conceptualization includes the main processes that influence syntactic complexity of performance. Constructing the complexity of performance appears to be the first actual step in producing language. Therefore, a speaker's first attempt, and consequently attention in the production process, tends to be directed to complexity. As a result, this notion of primary attention to complexity would be in contrast with the previous findings of research in SLA (VanPatten, 1990, 1994), which report that attentional resources are directed at elements of meaning and not form. In fact, the results of the current research show how demands of a task, in terms of its syntactic complexity requirement, or the condition and purpose of performing a task can primarily direct the participant's attention towards elements of form. It is worth mentioning that the

existing literature (VanPatten, 1990, 1994) that reports a contrasting view to the findings of the current research has looked at language performance in an instructional setting. Hence, it could be argued that this contrast exists because the present research has been conducted in an assessment setting, which might have influenced the participants' intentions to change their usual allocation of attentional resources and prioritize form, i.e. complexity and accuracy, over meaning.

Interestingly, the results of Study One showed that accuracy and fluency have greatly benefited from task structure as predicted by the hypotheses of the study. The inherent structure of a task, in fact, reduced the tension between the lexicalization and syntax processes and resulted in more accurate and fluent language performance. However, with complexity there was no straightforward effect that resulted from inherent task structure. The results from Study One clearly indicated that, based on Levelt's production model, complexity is influenced by other task characteristics or produced through different processes. But, what influenced complexity as a reflection of the processes occurring in the conceptualization stage remained unknown in Study One. In search of an answer to this question, Study Two was designed, hypothesizing type of grounding as the prime influence on complexity. The results of Study Two provide a clearer portrayal of complexity and the processes involved in the conceptualization level.

According to Levelt's model, conceptualization level is a message level where ideas are put together and a message is constructed. To influence the construction of the message at this level, therefore, a conceptual element should be presented to the speaker. It appears that foreground and background information is a conceptual characteristic of a task which is clearly capable of attracting and holding a speaker's attention more consistently. When the background information is incorporated into

the message at this message level, it will provide the speaker with a richer context to talk about. Adding background information to a picture story appears to be equal to increasing the number of propositions in a narrative. This richer context, and probably more propositions, would inspire the speaker at the conceptualization level. Then, there is a tendency in the speaker to utilize more complex language, i.e. more subordination, in the formulation phase to explain the rich context of the message level. However, this urgency of employing syntactic complexity hinders the formulation of the message. It could be argued that background information collocates with the macro-planning processes such as collecting and arranging information and will in turn influence the micro-planning processes of assigning the right forms to different chunks of information. Regarding the lexicalization processes, the results of Study Two indicate that the lexical processes involved in performance may not be an identical reflection of the processes involved in conceptualizing complexity[2]. However, to test such a hypothesis, more systematic research is required.

---

[2]As mentioned before, *vocd* measure of lexical variety was employed as a second measure of complexity in Study Two. The results of the analyses on *vocd* showed that lexical variety appeared to be a separate construct from complexity. However, it had associations with accuracy measures for the foreground tasks.

CHAPTER XI

Conclusions, Implications and Suggestions for Further Research

11.1 Introduction

This last chapter begins with a summary of the major findings of Study One and Study Two. It then discusses the conclusions and implications the results of the two studies would have for the fields of SLA and LT. The chapter will conclude with suggesting some potential areas for further research.

11.2 Conclusions from Study One

The results of the investigations of the data and the statistical analyses from Study One clearly indicate that the presence of task structure would reduce the cognitive load of oral narrative tasks and provide the test-takers with an opportunity to focus on fluency and accuracy. In effect, the results of Study One show that the presence of task structure improves accuracy and fluency of language performance of Iranian English language learners in an assessment setting. Performance on structured tasks, both problem-solution and schematic sequential, is progressively more accurate and more fluent than performance on unstructured tasks. Furthermore, tasks with a problem-solution structure have elicited the most fluent performances. However, the

results have indicated that task structure does not directly influence complexity of language performance.

The results of Study One further reveal that pre-task planning influences language performance in terms of fluency, accuracy and complexity. Planned performances are more accurate, fluent and complex than unplanned performances. The effects of pre-task planning on some measures of fluency are even greater than the effects of language proficiency. Interestingly, this suggests that better performance can be achieved if tasks and assessment conditions allow for planning compared to simply having a higher proficiency level.

More significantly, the results of the analyses on retrospective questionnaires show that task structure influences test-takers' perceptions of task difficulty. In effect, test-takers have perceived unstructured tasks as more difficult than structured tasks under both planned and unplanned conditions. Furthermore, non-planners have rated the tasks as more difficult than the planners have, i.e. non-planners have generally found the tasks more difficult to perform.

11.3 Conclusions from Study Two

The results from Study Two reveal very similar findings about task structure, i.e. presence of task structure in oral narrative tasks would influence language performance on the tasks. In Study Two, similar to the results of Study One, performances elicited by structured tasks are more fluent and accurate than performances elicited by unstructured tasks. Yet, complexity measures were not directly influenced by the structure of the tasks.

The results of the second study also show that grounding, i.e. providing foreground versus foreground and background information, is another significant task

characteristic that influences the performance of Iranian English language learners. Narrative tasks which contain foreground and background information elicit performances with greater syntactic complexity, while performance on tasks which contain only foreground information has been statistically less syntactically complex. Furthermore, performance on foreground information tasks is more fluent than performance on foreground and background information tasks. Grounding, however, does not influence accuracy in language performance on oral narrative tasks.

The results from both studies clearly throw light on a cognitive approach to task-based research, suggesting that as attentional resources available to L2 learners are limited, learners can only attend to some aspects of their performance while performing the cognitively demanding tasks. If a task, for instance, needs a lot of attention because it lacks structure or because there is some background information which is incorporated into the foreground events, there will be less attention available to be devoted equally well to different aspects of language performance. These findings clearly confirm the scheme Skehan (1996a, 1998) has proposed for analyzing task characteristics. As the results of both studies indicate, cognitive complexity and communicative stress are significant task characteristics and performance conditions which affect the difficulty level of the tasks and language performance on the tasks.

Another significant finding of both studies relates to the issue of tradeoffs between the three aspects of performance (Robinson, 2001; Skehan, 1998). The results show that there is a primary tradeoff between fluency and complexity of language performance, i.e. test-takers' attempt to produce more complex language would result in less fluency in the performance on the tasks. In addition, another tradeoff is seen between accuracy and complexity of performance on foreground tasks, which means that when the background information is not available accuracy and complexity interact. These

two tradeoff relationships confirm the findings of Skehan and Foster (1997) and Foster and Skehan (1996). However, the results of the factor analyses strongly indicate that presence of background information promotes accuracy and complexity. In other words, performance in tasks which contain both foreground and background information tends to be both more complex and more accurate, which confirms the findings of Robinson (2000). More significantly, these results suggest that the priority language learners and test-takers give to one aspect of language rather than to others appears to be a function of both the purpose of performing the task and the inherent characteristics of tasks. In effect, the common belief of prioritizing fluency of the message over its form might not apply to an assessment setting. For instance, with a task that includes foreground and background information, test-takers are more likely to give priority to complexity, whereas with a foreground task fluency would receive the primary attention. In addition, the priorities of a test-taker in an assessment setting may vary from the priorities of a language learner in classroom communication. More research is undoubtedly required to explore how, when and why test-takers attend to one aspect of language rather than the other(s).

A significant feature of both studies reported here has been employing a wide range of fluency measures to assess the fluency of the participants. The results of the various analyses on the fluency measures have revealed the multi-faceted nature of fluency and have pointed out the need for more research to provide a clearer picture of the construct of fluency. Temporal aspects of fluency and the length and density of utterances are mostly influenced by task characteristics and conditions. However, variations in repair fluency measures do not appear to follow a clear and predictable pattern. As reflected by statistical analyses in both studies, repair fluency measures seem to be representing a rather different aspect of fluency. The analyses of the

pausing patterns have also indicated that long pauses typically occurring in L2 learners' performance mainly happen in the middle of clauses rather than at clause boundaries. Furthermore, as the correlation between speech rate and number of pauses have shown it is evident that the mid-clause pauses, rather than the end-clause pauses, have a significant role in making second language performance less fluent.

11.4 Implications for SLA Research

The findings of this research strongly support the results of the previous research in task-based instruction on task difficulty (Skehan & Foster, 1996; Mehnert, 1998; Robinson, 2000), indicating that task characteristics and performance conditions influence task difficulty, performance on tasks and English language test-takers' perceptions of task difficulty. These findings imply that task characteristics should be considered in the selection and grading of tasks for both instructional and syllabus-design purposes. Tailoring specific characteristics of a task, i.e. task structure and grounding, among many others, would help language teachers achieve certain instructional goals such as improving fluency, accuracy or complexity of L2 learners' performance. This has implications for language pedagogy. In fact, in many language teaching classrooms employing tasks of appropriate characteristics, which are highly engaging and can push learners to generate more accurate and more complex performance, is a necessity. The interaction between different task characteristics should also be taken into consideration by SLA research, since this interactional effect clearly impacts on the complexity of selecting and grading tasks for pedagogic purposes.

Significant effects of pre-task planning would remind language teachers of the importance of providing language learners with some time to prepare if a more

reliable performance is targeted. The interesting results obtained from the comparison of the effect of pre-task planning time and language proficiency, i.e. that providing pre-task planning would enhance language performance more effectively than having a higher language proficiency level, has further potential pedagogic implications. This finding clearly suggests that lack of pre-task planning time could prevent learners from careful utilization of their true language ability, which would subsequently affect the learners' performance and teachers' judgement.

An important finding of the two studies for SLA is the allocation of attentional resources to different aspect of language performance. The results have clearly shown how language learners, while using language for communication, prioritize one aspect of performance over the other(s). The common belief, so far, has been that language learners would prioritize meaning over form. However, the findings of the two studies reported here indicate that task characteristics and performance conditions have a great role in channeling this prioritization. Therefore, by selecting tasks of suitable characteristics and by providing appropriate performance conditions, SLA researchers and language teachers would be able to channel the learners' attention to a specific direction and to encourage them to pay more attention to certain aspects of performance.

11.5 Implications for LT Research

The findings of this research have greater implications for LT research. First of all, unlike the results obtained from some studies (Elder et al., 2002; Iwashita et al., 2001), these results clearly demonstrate that task characteristics and conditions influence language performance in an assessment setting. Furthermore, it is now clear that cognitive complexity of 'task' influences task difficulty, which in turn impacts on

language performance. This suggests that in the process of designing tasks as assessment instruments, it is vital to pay considerable attention to very precise parameters of the task. The findings of the two studies reported here further indicate that task difficulty could reside in the task as a function of task characteristics. However, this does not counteract the possibility that task difficulty could be a function of the interaction between the task and the test-taker. To investigate if task difficulty is relative to any given test-taker, more research would necessarily be required.

Above all, the two studies reported here have contributed results which clarify the functioning of Skehan's model of oral language performance (Chapter III), and would help us to take the model beyond its schematic value and towards an empirical basis.

A crucial implication of the results of Study One and Study Two for LT relates to any interpretations and decisions that are made based on the test results. As task difficulty influences language performance, test results obtained from performance on tasks would not reflect only the language ability. In other words, since the performance is elicited by tasks which vary in the types of the language they require, this task variability may well introduce error into the assessment of the oral ability. In effect, the language performance on oral narrative tasks, at least to some extent, represents test-takers' language ability plus the effects that task characteristics and conditions have had on their performance. Without knowledge of these effects, the problem is that test scores that are assigned to test-takers might be artifactual, and would also be difficult to compare with results obtained under different conditions. For this reason, it is crucial to identify the detailed effects of task characteristics on different aspects of language performance.

As indicated earlier, there are occasions where a higher performance, i.e. increase in some aspects of fluency, is achieved because the test-takers are given more favorable performance conditions, i.e. pre-task planning time. This increase in some aspects of fluency suggests that the test scores assigned to test-takers may not reflect simply proficiency level, but the conditions under which a task is performed. Such test results are widely used and on the basis of these results important decisions with academic, social and professional values are usually made. Evidently, a slight variation in performance conditions or task characteristics would eventually impact on those decisions.

The results of both studies also suggest that more experimental studies will contribute to language test validation. Although some studies (Iwashita et al. 2001) have shown that some task characteristics, e.g. adequacy, immediacy and perspective, do not have significant effects on performance, the results of this research show that there are certain characteristics that directly impact on performance. The two studies reported here were able to investigate some task characteristics and conditions and the interaction between them that affect the language performance on tasks. However, there would be other characteristics and/or conditions that could inadvertently affect performance on tasks. By employing tasks of unknown characteristics and probable intrusive influences on performance, validity and reliability of the tests used to assess language performance of millions of test-takers every year should be scrutinized.

Another significant contribution of the two studies reported here for LT is clarification of the notions of task difficulty. As discussed in Chapter III, Brown et al (2002) regard task difficulty as a joint function of ability requirements and task characteristics. In other words, by task difficulty, they are trying to find out what *difficulty level* in a task a test-taker of certain ability could complete. In this way, in

an assessment setting, one could give test-takers of different proficiency levels tasks of appropriate difficulty so that they can cope with them. But the results of this research have confirmed that performance is multi-dimensional; i.e. fluency, accuracy and complexity appear to be the three independent areas of performance that can act in different patterns. Hence, test-takers are able to use their language ability and skills to compensate for certain aspects of language performance. Therefore, it is difficult to propose that a central criterion could be used to identify *difficulty level* of the tasks. Task characteristics may vary and this variation may connect systematically with different aspects of performance, but the problem is that these different aspects of performance may not function in unison; i.e. increasing one aspect of performance may not be associated with an increase in other aspects.

Last but not the least, these findings emphasize the importance of using tasks with large numbers of test-takers before they are employed for real assessment purposes as the weight of such large scale research will enhance confidence in the claims made by the research.

11.6 Limitations of This Research

The two research studies reported here were aimed at investigating the effects of task structure, grounding, pre-task planning time and language proficiency on language performance and perceptions of task difficulty of two groups of Iranian language learners of English. The results obtained from both studies have made identifiable contributions to the current understanding of tasks in TBI and TBA and are valuable in terms of implications they have for second language teaching and testing. However, certain limitations are exposed in this research. First, because of the reasons explained in Chapter One this research has employed a quantitative approach

to investigating the effects of task characteristics on language performance. It is certain that a qualitative approach to researching task characteristics and task difficulty would shed light on the complex relationships that exist among task difficulty, task characteristics, and test-taker characteristics and perceptions, and would have invaluable contributions to task-based research. Second, for reasons of scope and focus, the two studies reported here have investigated oral narrative tasks and explored only two characteristics of these tasks. More research studies are undoubtedly required to probe into other types of tasks and/or a variety of task characteristics. Although learner perceptions of task difficulty were investigated in Study One, for reasons of time and research design, they were not further investigated in Study Two. Further investigations of learner perceptions of task difficulty would broaden our perspective on how language learners feel about and perceive tasks and TBA. Another limitation of this research, which was caused by the practical restrictions of access, regards the participants of the studies. In this research, in effect, the participants are all L2 learners and from only one language background. Research studies that employ participants of different language backgrounds and native speakers of English would hypothetically provide results of higher generalizability. The last limitation of this research to be discussed here is its method used to assess language performance. In both studies in the current research an analytic detailed measures approach is adopted to assess the learners' performance on tasks. However, it is likely that employing additional methods of assessing language performance, e.g. using rating scales, would provide a more comprehensive evaluation of both the language performance and the effects of task characteristics on the learners' performance.

11.7 Suggestions for Further Research

The findings of this research show that there is a strong need for SLA and LT research to investigate task characteristics, task difficulty and the way they influence language performance. Systematic research that focuses on establishing a hierarchy of task difficulty will make a salient contribution to TBI and TBA. In the present context of language assessment, research programs are needed to explore how the range of characteristics and conditions of different tasks impacts upon language performance and test results. In addition, more research is required to investigate how different characteristics of tasks interact with one another and are, in turn, affected by performance conditions to influence task difficulty. In the present research, only one type of task which is typically used by international testing organizations, i.e. oral narrative tasks, has been examined. However, it is necessary to investigate other types of oral tasks that elicit dialogic and interactive performances.

To date, the majority of studies of task characteristics and language performance have been carried out with L2 learners without referring to a comparative study of native speakers. It seems important to know how native speakers of English would perceive task difficulty and perform different tasks. This kind of data would provide a basis for investigating the influence of tasks on different groups of speakers of English. It may also affect our understanding of the differences, if any, between the cognitive functioning of L1 and L2 speakers. Hence, another potential area of research is to investigate whether task characteristics have any effects on native speakers.

As the discussions of fluency measures in both studies suggested, no regular pattern resulting from the effects of task characteristics was discovered for the repair fluency measures. In effect, although task structure, grounding and planning time greatly influenced temporal measures of fluency, a similar influence was not observed for

reformulations, repetitions and replacements. Thus, more investigations are required to explore different aspects of repair fluency measures and further to explore what factors would have an impact on them.

Both studies in the present research adopted an analytic detailed measure approach to assessing oral language performance. A major issue to be dealt with in future LT research, therefore, is to investigate the differences emerging from the two methods of evaluating language performance, i.e. rating procedures and analytic detailed measures. As discussed earlier, due to practical restrictions of LT contexts, language performance is usually assessed through rating procedures. In SLA research, in contrast, a set of detailed measures is often used to assess language performance on tasks. Hence, a study that could employ both approaches to evaluating language performance would be able to compare the results obtained from the two methods and provide LT literature with a clearer picture of this debate.

Test-taker's perceptions of task difficulty have rarely been investigated in TBA. As the retrospective questionnaires in Study One showed, test-takers' perceptions of task difficulty were in line with the predicted cognitive complexity of tasks. However, detailed interviews with the test-takers are needed to provide LT and SLA research with a full account of how and why they find some tasks more difficult than others.

These are the main theoretical and experimental issues this research has raised and they are of prime significance as they will help language teachers, language test developers, syllabus designers and SLA and LT researchers to better understand the complexities involved in specific instances of language use. This type of information can throw light on some of the factors impacting on learner performance and it can help language educators to produce more effective teaching and particularly more appropriate testing.

REFERENCES

Alderson, J. C. (1991). Language testing in the 1990s: How far have we got? In S. Anivan (Ed.), *Current developments in language testing* (Vol. 25, pp. 1-26). Singapore: SEAMEO Regional Language Centre.

Alderson, J. C., & Banerjee, J. (2003a). State-of-the-art review: Language testing and assessment, part I. *Language Teaching, 34*, 213-236.

Alderson, J. C., & Banerjee, J. (2003b). State-of-the-art review: Language testing and assessment, part II. *Language Teaching, 35*, 79-113.

Allan, D. (1992). *Oxford Placement Test 2*. Oxford: Oxford University Press.

Allsop, J. (1996). *Elementary picture composition book*. London: Penguin.

Appel, G., & Lantolf, J. (1994). Speaking as mediation: A study of L1 and L2 text recall tasks. *The Modern Language Journal, 78*, 437-452.

Aston, G. (1986). Troubleshooting interaction with learners: The more the merrier? *Applied Linguistics, 7*, 128-143.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453-476.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bailey, K. M. (1985). If I hadn't known what I know now: Performance testing of foreign teaching assistants. In P. C. Hauptman, R. LeBlanc, & M. B.

Wesche (Eds.), *Second language performance testing* (pp. 153-180). Ottawa: University of Ottawa Press.

Baker, D. (1990). *A guide to language testing*. London: Arnold.

Bardovi-Harlig, K. (1998). Narrative structure and lexical aspect: Conspiring factors in second language acquisition of tense-aspect morphology. *Studies in Second Language Acquisition, 20*(4), 471-508.

Barkhuizen, G. P. (1998). Discovering learners' perceptions of ESL classroom teaching/learning activities in South African context. *TESOL Quarterly, 32*(1), 85-108.

Breen, M. (1984). Process syllabuses for the language classrooms. In C. Brumfit (Ed.), *General English syllabus design* (pp. 47-60). Oxford: Pergamon.

Breen, M. P. (1987). Learner contribution to task design. In N. C. Candlin & D. Murphy (Eds.), *Language learning tasks* (pp. 121-145). Englewood, Cliffs: Prentice Hall.

Brindley, G. (1987). Factors affecting task difficulty. In D. Nunan (Ed.), *Guidelines for the development of curriculum resources*. Adelaide: National Curriculum Resource Centre.

Brindley, G. (1994). Task-centered assessment in language learning: The promise and the challenge. In N. Bird, P. Falvey, A. Tsui, D. Allison, & A. McNeil (Eds.), *Language and language learning: papers presented at the Annual International language in Education Conference* (pp. 73-94). Hong Kong: Hong Kong education Department.

Brindley, G. (2001). Outcomes-based assessment in practice: Some examples and emerging insights. *Language Testing, 18*(4), 393-407.

Brock, C. (1986). The effects of referential questions on ESL classroom discourse. *TESOL Quarterly, 20*(1), 47-59.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1-15.

Brown, J. D., Hudson, T. D., Norris, J. M., & Bonk, W. (2002). *An investigation of second language task-based performance assessment*. Honolulu: University of Hawaii Press.

Brumfit, C., & Johnson, K. (1979). *The Communicative Approach to Language Teaching*. Oxford: Oxford University Press.

Bygate, M. (1996). Effects of task repetition: Appraising the developing language of learners. In D. Willis, J. (Ed.), *Challenge and change in language teaching*. London: Heinemann.

Bygate, M. (1999). Quality of language and purpose of task: Patterns of learners' language on two oral communication tasks. *Language Teaching Research, 3*(3), 185-214.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing*. London: Longman.

Bygate, M., Skehan, P., & Swain, M. (2001). *Researching pedagogic tasks: Second language learning, teaching and testing*. London: Longman.

Canadian Language Benchmarks for Adult ESL. (2000). Canadian Language Centre. *www.language.ca/Main_En/Benchmarks*

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards, & R. W. Schmidt (Eds.), *Language and communication* (pp. 2-27). London: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Candlin, C. (1987). Towards task-based language learning. In C. M. Candlin, & D. Murphy (Eds.), *Language learning tasks*. London: Prentice Hall.

Candlin, C. (1984). Syllabus design as a critical process. In C. Brumfit (Ed.), *General English syllabus design* (pp. 29-46). Oxford: Pergamon.

Carrell, P. (1985). Facilitating ESL reading by teaching text structure. *TESOL Quarterly, 19*(4), 727-749.

Carroll, J. B. (1968). The psychology of language testing. In A. Davies (Ed.), *Language testing symposium: A psycholinguistic perspective* (pp. 46-69). London: Oxford University Press.

Carroll, J. B. (1985). LT + 25, and beyond? Comments. *Language Testing, 3*(2), 123-129.

Carroll, J. B. (1987). New perspectives in the analysis of abilities. In R. Ronning, J. A. Glover, J. C. Conoely, & J. C. Witt (Eds.), *The influence of cognitive society on testing* (pp. 267-284). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chalhoub-Deville. (2001). Task-based assessments: Characteristics and validity evidence. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language teaching, learning and testing*. London: Longman.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing, 14*(1), 3-22.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369-383.

Chappelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-71). Cambridge: Cambridge University Press.

Chaudron, C. (1988). *Second language classrooms: Research on teaching and learning*. Cambridge: Cambridge University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press.

Clark, J. L. D. (1975). Theoretical and technical considerations in oral proficiency testing. In R. L. Jones & B. Spolsky (Eds.), *Testing language proficiency* (pp. 10-24). Arlington, Va.: Centre for Applied Linguistics.

Coady, J. (1979). A psycholinguistic model of the ESL reader. In R. Macky, B. Barkman, & R. R. Jordan (Eds.), *Reading in a second language: Hypotheses, organisation and practice* (pp. 110-138). Rowley, MA: Newbury House.

Cohen, A. (1984). On taking language tests: What are the students report? *Language Testing, 1*, 70-81.

Cohen, A. (1996). Verbal reports as a source of insights into second language learning strategies. *Applied Language Learning, 7*, 5-24.

Coughlan, P., & Duff, P. (1994). Same task, different activities: Analysis of a second language acquisition task from an activity theory perspective. In J.

Lantolf & G. Appel (Eds.), *Vygotskian approaches to second language research* (pp. 173-195). Norwood, NJ: Ablex.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition, 11*, 367-383.

Davies, A. (1978). Language testing: Survey articles 1 and 2. *Language Teaching and Linguistics Abstracts, 11*, 145-159.

Davies, A. (2003). Three heresies of language testing research. *Language Testing, 20*(4), 355-368.

Dotano, R. (1994). Collective scaffolding in second language learning. In J. Lantolf & G. Appel (Eds.), *Vygotskian approach to second language research* (pp. 33-56). Norwood, NJ: Ablex.

Doughty, C. (1997). *Meeting the criteria of focus on form.* Paper presented at the Second Language Research Forum, Michigan State University.

Doughty, C. (2001). Cognitive understanding of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 206-258). Cambridge: Cambridge University Press.

Doughty, C., & Willaims, J. (1998). Pedagogical choices in focus on form. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 197-262). New York: Cambridge University Press.

Dry, H. (1983). The movement of narrative time. *Journal of Literary Semantics, 12*, 19-53.

Duff, P. (1986). Another look at interlanguage talk: Talking task to task. In R. Day (Ed.), *Talking to learn.* Rowley, Mass.: Newbury House.

Duff, P. (1993). Tasks and interlanguage performance: A SLA perspective. In G. Crookes & S. Gass (Eds.), *Tasks and language learning: Integrating theory and practice*. Clevedon, Avon: Multilingual Matters.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer. *Language Testing, 19*(4), 343-368.

Ellis, N. (1996). Sequencing in SLA: Phonological memory, chunking and points of order. *Studies in Second Language Acquisition, 18*(1), 91-126.

Ellis, R. (1985). *Understanding second language acquisition*. Oxford: oxford University Press.

Ellis, R. (1987). Interlanguage variability in narrative discourse: Style shifting in the use of the past tense. *Studies in Second Language Acquisition, 9*, 12-20.

Ellis, R. (2000). Task-based research and language pedagogy. *Language Teaching Research, 4*(3), 193-220.

Ellis, R. (2003). *Task-based language teaching and testing*. Oxford: Oxford University Press.

Farhady, H., & Abbasian, G. R. (1999). Test method, Level of language proficiency and underlying structure of language ability. *Journal of Humanities Alzahra University, 9*(29), 58-84.

Fillmore, C. J. (1979). On fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 43-61). Ann Arbor: University of Michigan Press.

Foster, P. (1998). A classroom perspective on the negotiation of meaning. *Applied Linguistics, 19*, 1-23.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performances. *Studies in Second Language Acquisition, 18*, 299-323.

Foster, P., & Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research, 3*(3), 215-247.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language. *Applied Linguistics, 21*(3), 354-375.

Fotos, S., & Ellis, R. (1991). Communicating about grammar: A task-based approach. *TESOL Quarterly, 25*, 605-628.

Freed, B. F. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243-266). Ann Arbor: University of Michigan Press.

Fulcher, G. (1997). The testing of L2 speaking. In C. Clapham & D. Corson (Eds.), *Language testing and assessment* (Vol. 7, pp. 75-85). Dordrecht: Kluwer Academic Publishers.

Fulcher, G. (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics, 20*(2), 221-236.

Fulcher, G., & Marquez Reiter, R. (2003). Task difficulty in speaking tests. *Language Testing, 20*(3), 321-344.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing, 19*(2), 133-167.

Gipps, C. (1994). *Beyond testing: Towards a theory of educational assessment*. London: Falmers Press.

Givon, T. (1985). Function, structure, and language acquisition. In D. Slobin (Ed.), *The crosslinguistic study of language acquisition: Vol. 1* (pp. 1008-1025). Hillsdale, NJ.: Erlbaum.

GoldWave. (2001). Version v4.25: *www.goldwave.com*.

Graham, S. J. (2004). Giving up on modern foreign languages? Students' perceptions of learning French. *Modern Language Journal, 88*(2), 171-198.

Hall, C. (1993). The direct testing of oral skills in university foreign language teaching. *IRAL, 31*(1), 23-38.

Harely, T. (2001). *The psychology of language*. East Sussex: Psychology Press LTD.

Harley, B., Allen, J. P. B., Cummins, J., & Swain, M. (1990). *The development of second language proficiency*. Cambridge: Cambridge University Press.

Hatch, E. (1978). *Second language acquisition: A book of readings*. Rowley, Mass.: Newbury House.

Hatch, E., & Farhady, H. (1984). *Research Design and statistics*. Tehran: Rahnama.

Heaton, J. B. (1966). *Composition through pictures*. Essex: Longman.

Henning, G. (1987). *A guide to language testing*. Cambridge, Mass.: Newbury House.

Henning, G. (1990). National issues in individual assessment: The consideration of specialization bias in university language screening tests. In J. H. L. d. Jong & D. K. Stevensen (Eds.), *individualizing the assessment of language abilities* (pp. 38-50). Clevedon, Avon: Multilingual Matters.

Hoey, M. (1983). *On the surface of discourse*. London: George Allan and Unwin.

Hooper, P. J., & Thompson, S. A. (1980). Transitivity in grammar and discourse. *Language, 56*, 251-299.

Hughes, A., & Porter, D. (1983). *Current developments in language testing.* London: Academic Press.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels: NCTE Research Report* (3). London: National Council of Teachers of English.

Hymes, D. (1971). Competence and performance in linguistic theory. In R. Huxley & E. Ingram (Eds.), *Language acquisition: Models and methods* (pp. 27-49). New York: Academic Press.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth: Penguin.

Iwashita, N., & Elder, C. (1997). Expert feedback? Assessing the role of test-taker reactions to a proficiency test for teachers of Japanese. *Melbourne Papers in Language Testing, 6*, 53-67.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning, 51*(3), 401-436.

Jones, L. (1980). *Notions in English.* Cambridge: Cambridge University Press.

Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31-36.

Kenyon, D. (1992). *Introductory remarks at symposium on development and use of rating scales in language testing.* Vancouver: Canada.

Kintsch, W., & Van Dijk, T. A. V. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394.

Kobayashi, M. (1995). *The effects of text structure and test format on the measurement of reading comprehension.* Unpublished Ph.D. Thesis, Thames Valley University, London.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19*(2), 193-220.

Kopenon, M., & Riggenbach, H. (2000). *Overview: Varying perspectives on Fluency*. Ann Arbor, Michigan: University of Michigan Press Robinson.

Krashen, S. D. (1980). *Second language acquisition and second language learning*. New York: Prentice Hall.

Krashen, S. D. (1985). *The input hypothesis*. London: Longman.

Krashen, S. D., & Terrell, T. D. (1983). *The natural Approach: Language acquisition in the classroom*. Oxford: Pergamon.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.

Lantolf, J. (2000). *Sociocultural theory and second language learning*. Oxford: Oxford University Press.

Lee, J. F., Cadierno, T., Glass, W. R., & Van Patten, B. (1997). The effects of lexical and grammatical cues on processing past temporal reference in second language input. *Applied Language Learning, 8*, 1-23.

Leeser, M. J. (2004). Learner proficiency and focus on form during collaborative dialogue. *Language Teaching Research, 8*(1), 55-81.

Leung, C. (2000). Teacher assessment and psychometric theory: A case of paradigm crossing. *Language Testing, 17*(2), 161-184.

Leung, C. (2001). Evaluation of content-language learning in the mainstream classroom. In B. Mohan, C. Leung, & C. Davison (Eds.), *English as a second language in the mainstream: teaching, learning and identity* (pp. 177-198). Harlow, Essex: Longman.

Leung, C., Harris, R., & Rampton, B. (2004). Living with inelegance in qualitative research on task-based learning. In K. Toohey & B. Norton (Eds.), *Critical pedagogies and language learning* (pp. 242-267). Cambridge: Cambridge University Press.

Leung, C., & Teasdale, A. (1997). Raters' understanding of rating scales as abstracted concept and as instruments for decision-making. *Melbourne Papers in Language Testing, 6*, 45-70.

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W. (1993). Language use in normal speakers and its disorder. In G. Blanken, H. Dittman, H. Grimm, J. Marshal, & C. Wallesch (Eds.), *Linguistic disorders and pathologies* (pp. 1-15). Berlin: de Gruyter.

Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing, 17*(1), 43-64.

Lightbown, P., & Spada, N. (1999). *How languages are learned*. Oxford: Oxford University Press.

Littlewood, W. (1992). *Teaching oral communication: A methodological framework*. Oxford: Blackwell.

Littlewood, W. (1999). Second language teaching methods. In B. Spolsky (Ed.), *Concise encyclopedia of educational linguistics* (pp. 658-668). Oxford: Elsevier.

Long, M., Ingaki, S., & Ortega, L. (1998). The role of implicit negative feedback in SLA: Models and recasts in Japanese and Spanish. *Modern Language Journal, 82*, 357-371.

Long, M., & Norris, J. M. (2000). Task-based language teaching and assessment. In M. Byram (Ed.), *Encyclopedia of language teaching* (pp. 597-603). London: Routledge.

Long, M. H. (1983). Does second language instruction make a difference? A review of the research. *TESOL Quarterly, 17*, 359-382.

Long, M. H. (1985). A role for instruction in second language acquisition: Task-based language teaching. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (pp. 77-99). Clevedon, Avon: Multilingual Matters.

Long, M. H. (1988). Instructed interlanguage development. In L. Beebe (Ed.), *Issues in second language acquisition: Multiple perspectives* (pp. 115-141). New York: Newbury House.

Long, M. H. (1989). *Task, group, and task-group interactions.* Paper presented at the University of Hawaii working papers in ESL.

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. Coste, R. Ginsburg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspectives* (pp. 39-52). Amsterdam: Benjamins.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). San Diego, CA: Academic Press.

Long, M. H. (2000). Focus on form in task-based language teaching. In R. D. Lambert & E. Shohami (Eds.), *Language policy and language pedagogy* (pp. 179-192). Amsterdam: Benjamins.

Long, M. H., & Crookes, G. (1992). Three approaches to task-based syllabus design. *TESOL Quarterly, 26*, 27-56.

Long, M. H., & Robinson, P. (1998). Focus on form: Theory, research, and practice. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 14-41). New York: Cambridge University Press.

Loschky, L., & Bley-Vroman, R. (1993). Grammar and task-based methodology. In G. Crookes & S. Gass (Eds.), *Tasks in language learning: Integrating theory and practice* (pp. 123-167). Clevedon, Avon: Multilingual Matters

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Lyster, R. (1998). Recasts, repetition, and ambiguity in L2 classroom discourse. *Studies in Second Language Acquisition, 20*(1), 51-81.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classroom. *Studies in Second Language Acquisition, 19*(1), 37-66.

Mackey, A., Gass, S., & McDonough, K. (2000). Do learners recognize implicit negative feedback as feedback? *Studies in Second Language Acquisition, 22*(4), 471-497.

Malinowski, B. (1935). *Coral gardens and their magic*. London: Allen and Unwin.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85-104.

Mandle, H., & Levin, J. R. (1989). *Knowledge acquisition from text and pictures*. Amsterdam: North-Holland.

Mandler, J. M. (1978). A code in the node: The use of story schemata in retrieval. *Discourse Processes, 1*(1), 14-35.

Martin, J. R. (1985). *Factual writing: Exploring and challenging social reality.* Victoria: Deakin University Press.

McLaughlin, B., Rossman, T., & McLeod, B. (1983). Second language learning: An information-processing perspective. *Language Learning, 33*, 135-158.

McMillan, J. H. (1996). *Educational research: Fundamentals for consumers.* New York: Harper Collins.

McNamara, T. F. (1995). Modelling language performance: Opening Pandora's box. *Applied Linguistics, 16*(2), 159-175.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. F. (2000). *Language testing*. Oxford: Oxford University Press.

Mehnert, U. (1998). The effects of different length of time for planning on second language performance. *Studies in Second Language Acquisition, 20*(1), 83-108.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256.

Meyer, B. J. F., Brandt, D. V., & Bluth, G. J. (1980). Use of top-level structure in text: Key for reading comprehension of ninth-grade students. *Reading Research Quarterly, 16*(1), 72-103.

Mohan, B. A. (1986). *Language and content*. Reading, MA: Addison-Wesley.

Mohan, B. A. (1991). LEP students and the integration of language and content: Knowledge structures and tasks. In C. Simich-Dudgeon (Ed.), *Proceedings of the first symposium on limited English proficient students' issues*. Washington, DC: Office of Bilingual Education and Minority Language Affairs.

Mohan, B. A. (2001). The second language as a medium of learning. In B. Mohan, Leung. C., & C. Davison (Eds.), *English as a second language in the mainstream: Teaching, learning and identity* (pp. 107-126). Harlow, Essex: Longman.

Navon, D. (1989). The importance of being visible: On the role of attention in a mind viewed as an anarchic intelligence system. *European Journal of Cognitive Psychology, 1*, 191-238.

Neumann, O. (1996). Theories of attention. In O. Neumann & A. Sanders (Eds.), *Handbook of perception and action Vol. 3: Attention* (pp. 389-446). San Diego: CA: Academic Press.

Nicholas, H., Lightbown, P. M., & Spada, N. (2001). Recasts as feedback to language learners. *Language Learning, 51*(4), 719-758.

Norris, C. B. (1991). Evaluating English oral skills through the technique of writing as if speaking. *System, 19*(3), 203-216.

Norris, J.M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessment*. Honolulu: University of Hawaii Press.

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task-based second language performance assessment. *Language Testing, 19*(4), 395-418.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*, 217-528.

Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge: Cambridge University Press.

Nunnally, J. O. (1978). *Psychometric theory*. New York: McGraw-Hill.

O'Conner, S. (1989). *First Certificate in English*. Oxford: Oxford University Press.

Oller, J. W. J. (1976). Evidence of a general language proficiency factor: An expectancy grammar. *Die Neuren Sprachen, 76*, 165-174.

Oller, J. W. J. (1986). Communication theory and testing: What and how. In C. W. Stansfield (Ed.), *Towards communicative competence testing: Proceedings of the second TOEFL invitational conference* (pp. 104-155). Princeton, NJ: Educational Testing Service.

O'Loughlin, K., & Wigglesworth, G. (2003). *Task design in IELTS academic writing task 1: The effect of quantity and manner of presentation of information on candidate writing.* Paper presented at the LTRC, Reading, UK.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement*. London: Pinter Publishers.

Oppenheim, N. (2000). The importance of recurrent sequences for nonnative speaker fluency and cognition. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 220-241). Ann Arbor: University of Michigan Press.

Ortega, L. (1995). *Planning and second language oral performance.* Unpublished Unpublished MA thesis, University of Hawaii, Honolulu.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition, 21*, 109-148.

Ortega, L. (2003). Syntactic complexity measure and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492-518.

Pallant, J. (2001). *SPSS survival manual*. Buckingham: Open University Press.

Pica, T. (1994). Research on negotiation: What does it reveal about second language learning conditions, processes, and outcomes? *Language Learning, 44*, 493-527.

Pica, T. (1997). Second language teaching and research relationships: A North American View. *Language Teaching Research, 1*, 48-72.

Pica, T., & Doughty, C. (1985). Input and interaction in the communicative language classroom: A comparison of teacher-fronted and group activities. In S. Gass & C. Madden (Eds.), *Input and second language acquisition* (pp. 112-145). Rowley, Mass.: Newbury House.

Pica, T., Halliday, L., Lewis, N., & Morgenthaler, L. (1989). Comprehensible outputs as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition, 11*(1), 63-90.

Pienemann, M. (1984). Learnability and syllabus construction. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language performance* . Avon: Multilingual Matters.

Polanyi-Bowditch, L. (1976). Why the whats are when: Mutually contextualizing realms of narrative. *Berkeley Linguistics Society, 2*, 59-77.

Prabhu, N. S. (1987). *Second language pedagogy*. Oxford: Oxford University Press.

Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high and low ability test takers: A structural equation modeling approach. *Language Testing, 15*(3), 333-379.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Harlow: Longman.

Rahimpour, M. (1997). *Task condition, task complexity and variation in oral L2 discourse.* Unpublished doctoral dissertation, University of Queensland, Brisbane, Australia.

Raupach, M. (1980). Temporal variables in first and second language production. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech: Studies in honour of Freida Goldman-Eissler* (pp. 49-60). The Hauge: Mouton.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1-32.

Reed, D. J. (1992). The relationship between criterion-based levels of oral proficiency and norm--referenced scores of general proficiency in English as a second language. *System, 20*(3), 329-345.

Reinhart, T. (1984). Principles of gestalt perception in the temporal organization of narrative texts. *Linguistics, 22*, 779-809.

Richards, B. J. (1978). Type/token ratio: what do they really tell us? *Journal of Child Language*, 14, 201-9.

Richards, J. C. (1985). *The context of language teaching*. Cambridge: Cambridge University Press.

Richgels, D. J., McGee, L. A., Lomax, R. G., & Sheard, C. (1987). Awareness of text structures: Effects of a recall of expository text. *Reading Research Quarterly, 22*(2), 177-196.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning, 45*(1), 99-140.

Robinson, P. (1996). Learning simple and complex second language rules under implicit, incidental, rule-search, and instructed conditions. *Studies in Second Language Acquisition, 18*, 27-67.

Robinson, P. (2000). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27-55.

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287-318). Cambridge: Cambridge University Press.

Samuda, V. (2001). Guiding relationships between form and meaning during task performance: The role of the teacher. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks second language learning, teaching and testing* (pp. 119-141). London: Longman.

Savignon, S. J. (1983). *Communicative competence: Theory and classroom practice.* Reading, Mass.: Addison-Wesley.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding.* Hillsdale, N. J.: Earlbaum.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*, 129-158.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-33). Cambridge: Cambridge University Press.

Scott, M. L., & Madsen, H. S. (1983). The influence of retesting on test affect. In J.W. J. Oller (Ed.), *Issues in language testing research* (pp. 270-279). Rowley, Mass.: Newbury House.

Seliger, H., & Shohamy, E. (1989). *Second language research methods*. Oxford: Oxford University Press.

Shaw, M. L., & Shaw, P. (1978). A capacity allocation model for reaction time. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 596-598.

Shohamy, E. (1982). Affective considerations in language testing. *The Modern Language Journal, 66*, 13-17.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics, 15*, 118-221.

Shohamy, E. (2001). *The power of tests*. Harlow: Pearson Education Limited.

Skehan, P. (1996a). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38-62.

Skehan, P. (1996 b). Second language acquisition research and task-based instruction. In J. Willis & D. Willis (Eds.), *Challenge and change in language teaching* (pp. 17-31). Oxford: Heinemann.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swains (Eds.), *Researching pedagogic tasks, second language learning, teaching and testing* (pp. 167-185). London: Longman.

Skehan, P. (2003). Task-based instruction. *Language Teaching, 36*, 1-14.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*(3), 185-212.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49*(1), 93-120.

Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 183-205). Cambridge: Cambridge University Press.

Smith, B. (1982). *In your own words*. Kent: Hodder and Sloughton.

SoundScriber. (1998): *www.lsa.umich.edu.*

Spolsky, B. (1977). Language testing: Art or science. In G. Nickel (Ed.), *Proceedings of the fourth International Congress of Applied Linguistics* (Vol. 3, pp. 60-85). Stuttgart: Hochschulverlag.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing, 2*(1), 31-40.

Spolsky, B. (1995). *Measured words*. Oxford: Oxford University Press.

Stern, D. (1992). *American sound and style for all speakers of English as a Second Language*. Los Angeles: Video Language Program.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. Mahway, NJ.: Lawrence Erlbaum.

Swails, J. (1985). *Episodes in ESP*. Oxford: Pergamon.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235-253). Rowley, Mass.: Newbury House.

Swain, M. (1991). French immersion and its offshoots: Getting two for one. In B. Freed (Ed.), *Foreign language acquisition: Research and the classroom* (pp. 91-103). Lexington, MA: Heath.

Swain, M. (1993). Second-language testing and second-language acquisition: Is there a conflict with traditional psychometrics? *Language Testing, 10*(2), 191-207.

Swain, M., & Lapkin, S. (2001). Focus on form through collaborative dialogue: Exploring task effects. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks second language learning, teaching and testing* (pp. 99-119). London: Longman.

Swan, M. & Walter, C. (1990). *New Cambridge English Course: Workbook Level 1*, Cambridge: Cambridge University Press.

Tabachnic, B. G., & Fidell, L. S. (1996). *Using multivariate statistics*. New York: Harper Collins.

Tarone, E. (1979). Interlanguage as chameleon. *Language Learning, 29*, 181-192.

Thompson, I. (1995). A study of inter-rater reliability of the ACTFL Oral Proficiency Interview in five European languages: Data from ESL, French, German, Russian and Spanish. *Foreign Language Annals, 28*(3), 407-422.

Ting, S. C. C. (1996). Planning time, modality and second language task performance: Accuracy and fluency in the acquisition of Chinese as a second language. *The University of Queensland Working Papers in Language and Linguistics, 1*, 31-64.

Tomlin, R. (1984). The treatment of foreground-background information in the on-line descriptive discourse of second language learner. *Studies in Second Language Acquisition, 6*, 115-142.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*(1), 84-115.

Turner, A. (1992). *Patterns of thinking*. Sydney: Primary English Teaching Association.

Upshur, J. A. (1979). Functional proficiency theory and a research role for language tests. In J. E. Briere & F. B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp. 75-100). Washington DC: TESOL.

Ur, P. (1984). *Teaching listening comprehension*. Cambridge: Cambridge University Press.

Van Dijk, T. A. (1977). *Text and context: Explorations in the semantics and pragmatics of discourse*. London: Longman.

Van Dijk, A., & Kintsch, W. (1978). Cognitive psychology and discourse: Recalling and summarizing stories. In W. U. Dressler (Ed.), *Current trends in text linguistics* (pp. 78-105). Berlin: Walter de Gruyter.

Van Lier, L., & Matsu, N. (2000). Varieties of conversational experience: Looking for learning opportunities. *Applied Language Learning, 11*(2), 265-287.

VanPatten, B. (1990). Attending to form and content in the input: An experiment in consciousness. *Studies in Second Language Acquisition, 12*, 287-301.

VanPatten, B. (1994). Evaluating the role of consciousness in second language acquisition: Terms, linguistic features & research methodology. *AILA Review, 11*, 27-36.

von Stutterheim, C. (1991). Narrative and description: Temporal reference in second language acquisition. In C. A. Ferguson & T. Huebner (Eds.), *Crosscurrents in second language acquisition and linguistic theories* (pp. 385-403). Amsterdam: Benjamins.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA.: Harvard University Press.

Wesche, M. B. (1985). Introduction. In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing*. Ottawa: University of Ottawa Press.

Widdowson, H. G. (1978). *Teaching language as communication*. Oxford: Oxford University Press.

Widdowson, H. G. (1979). *Explorations in applied linguistics*. Oxford: Oxford University Press.

Widdowson, H. G. (1989). Knowledge of language and ability for use. *Applied Linguistics, 10*(2), 37-67.

Widdowson, H. G. (1990). *Aspects of language teaching*. Oxford: Oxford University Press.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 21-44.

Wigglesworth, G. (2001). Influences on performance in task-based oral assessments. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language teaching, learning and testing*. London: Longman.

Wilkins, D. (1976). *Notional syllabuses*. Oxford: Oxford University Press.

Wilkins, D., A. (1999). Second language teaching. In B. Spolsky (Ed.), *Concise encyclopedia of educational linguistics* (pp. 656-658). Oxford: Cambridge University Press.

Willis, J. (1996). *A framework for task-based learning*. London: Longman.

Willis, J., & Willis, D. (1996). *Challenge and change in language teaching*. London: Heinemann.

Winn, W. (1991). Learning from maps and diagrams. *Educational Psychology Review, 3*, 211-247.

Winter, E. O. (1976). *Fundamentals of information structure: A pilot manual for further development according to student need*s. Hatfield: The Hatfield Polytechnic, Linguistics group, School of humanities.

Wood, D., Bruner, J., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*, 89-100.

Young, R. (1995). Conversation style in language proficiency interview. *Language Learning, 45*(1), 3-45.

Young, R. F. (2000). *Interactional competence: Challenges for validity.* Paper presented at the 21st LTRC, Vancouver, Canada.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics, 24*(1), 1-27.

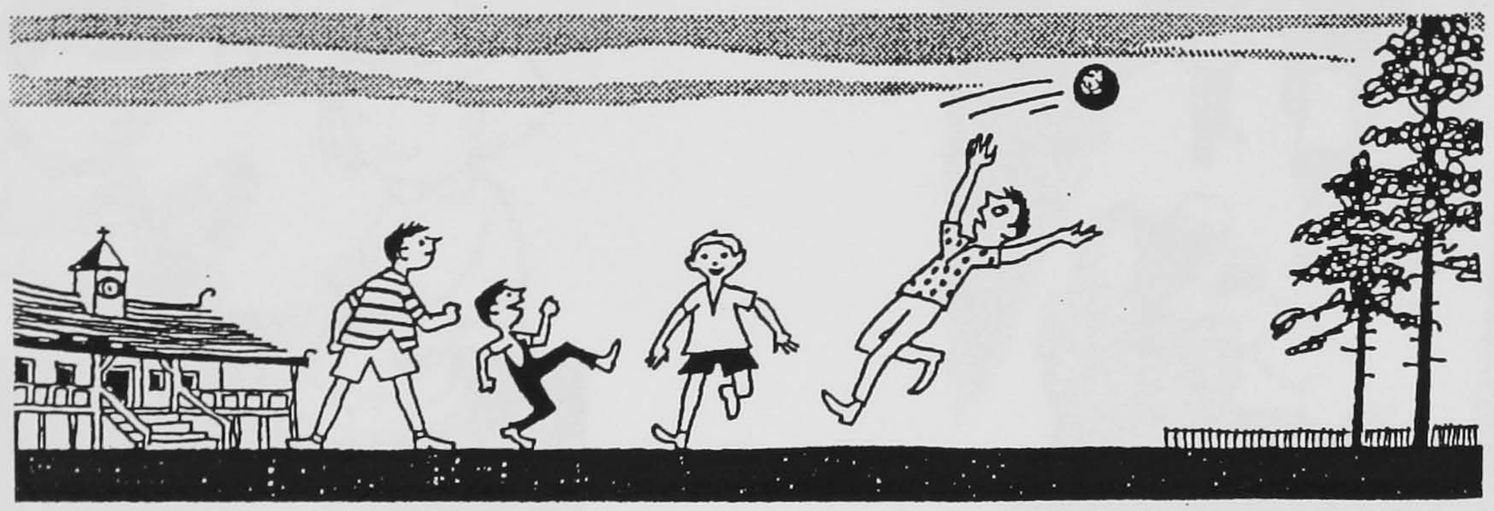
Zeinder, M. (1990). College students' reactions towards key facets of classroom testing. *Assessment and Evaluation in Higher Education, 15*, 85-106.

List of Figures

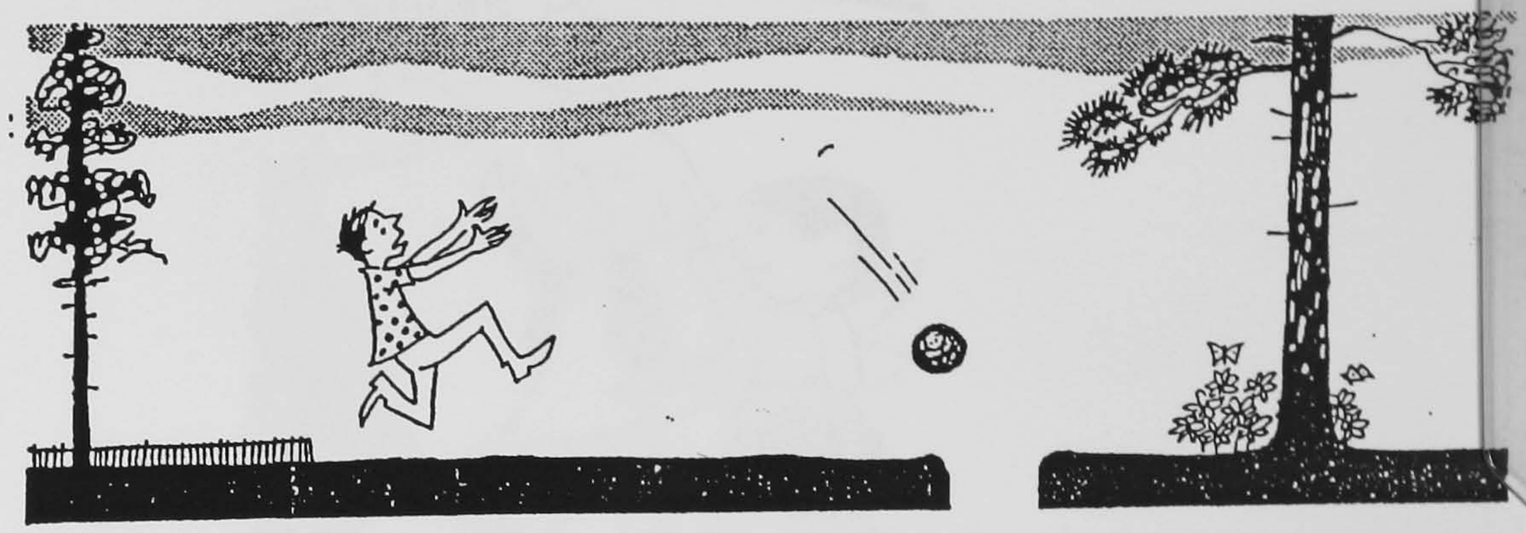
List of Tables

Appendix 1

Football



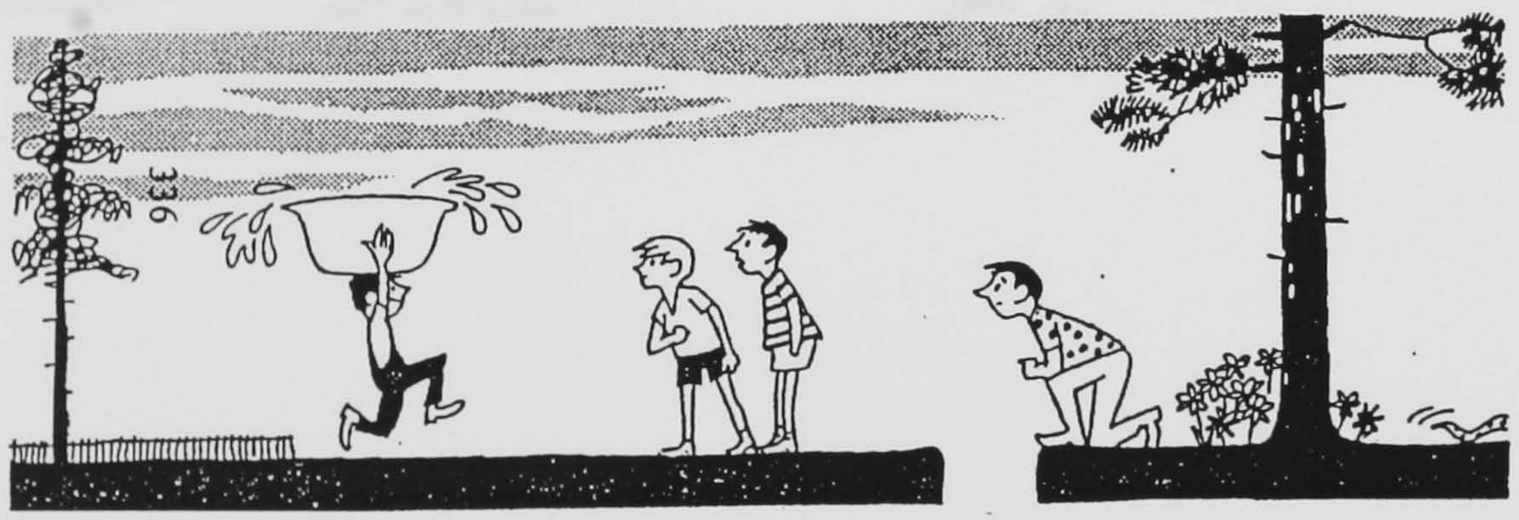
1



2

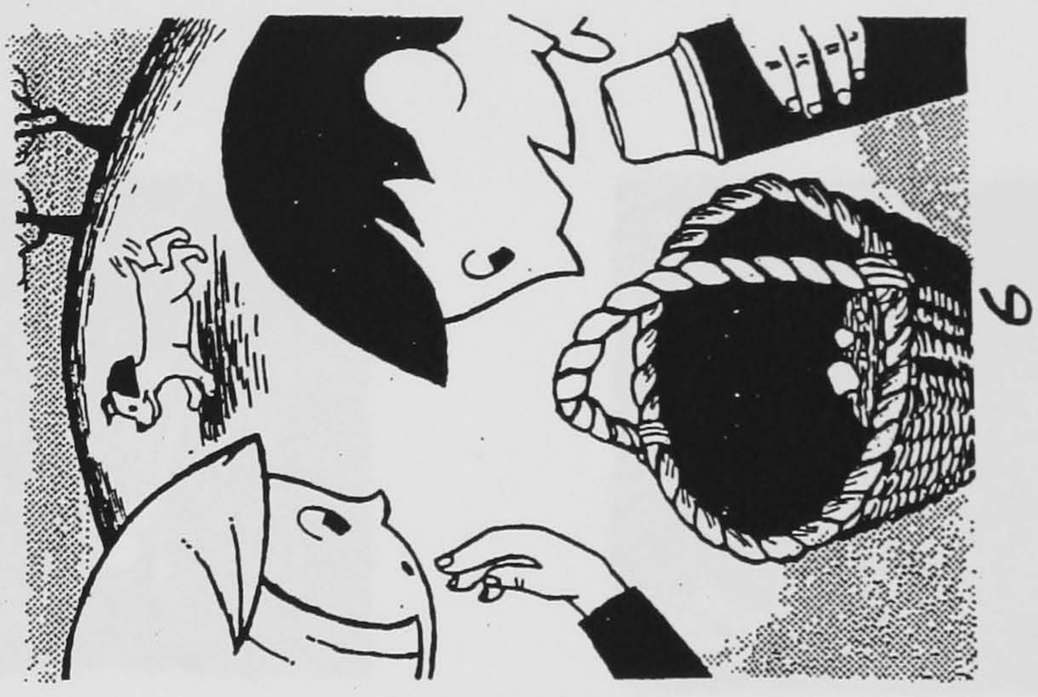
3

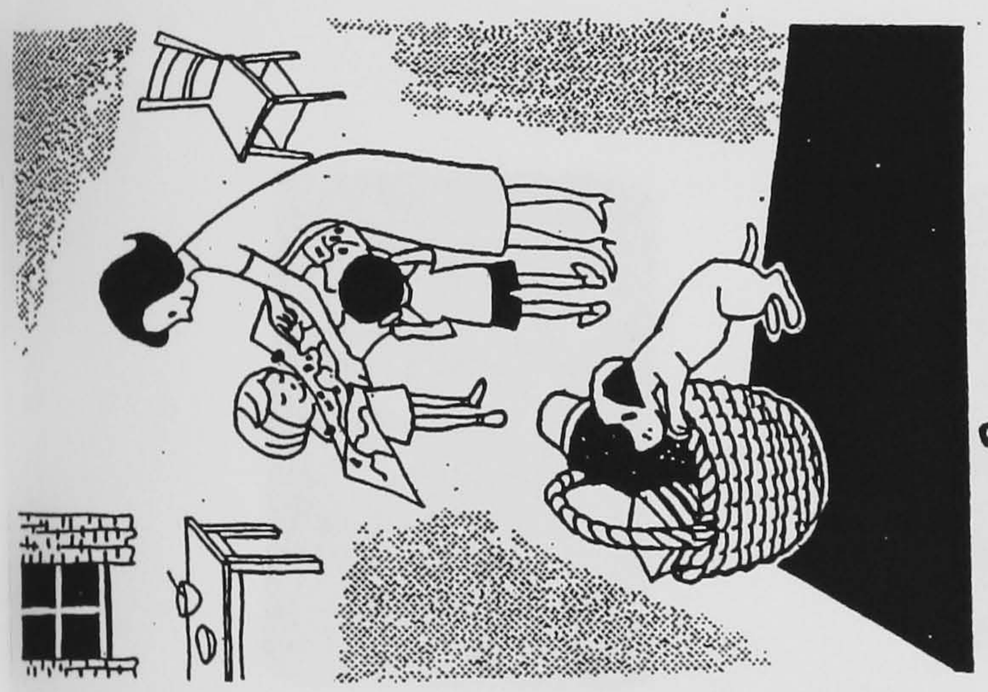
4

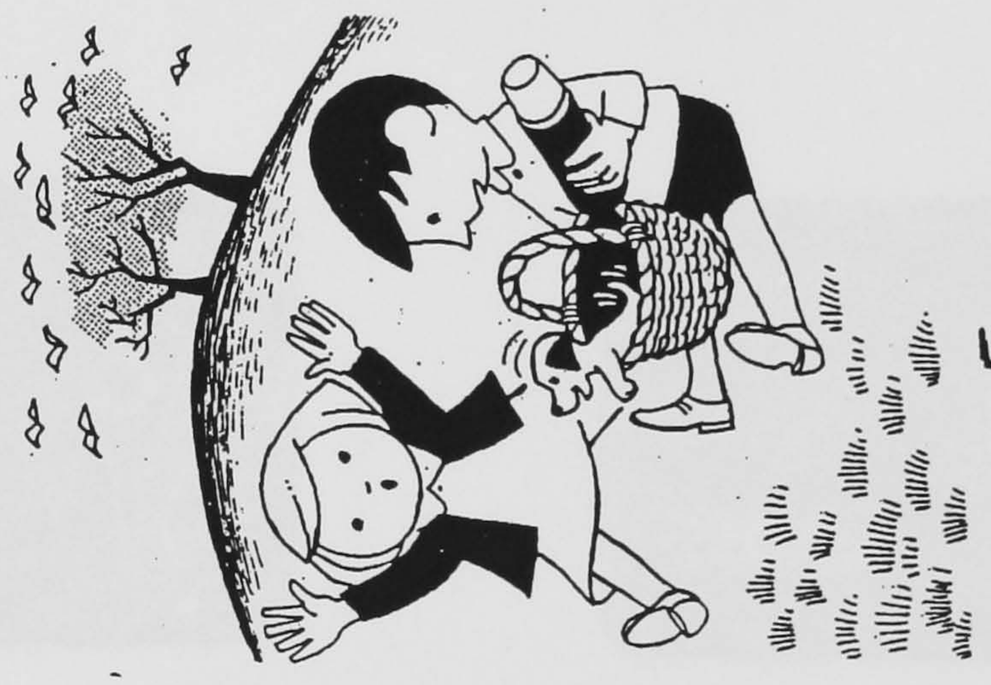


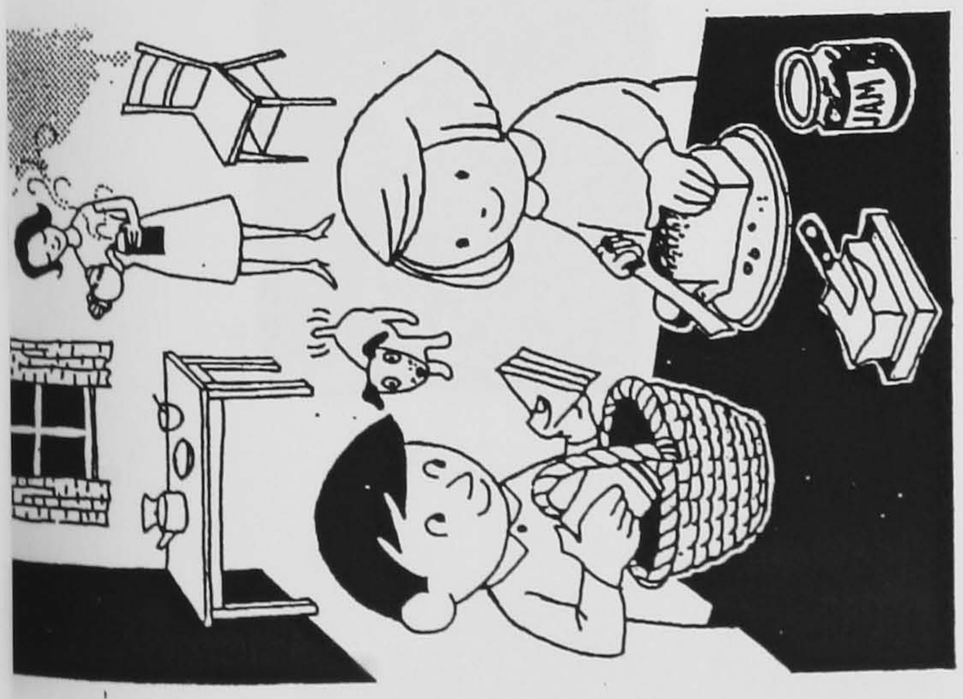
Picnic

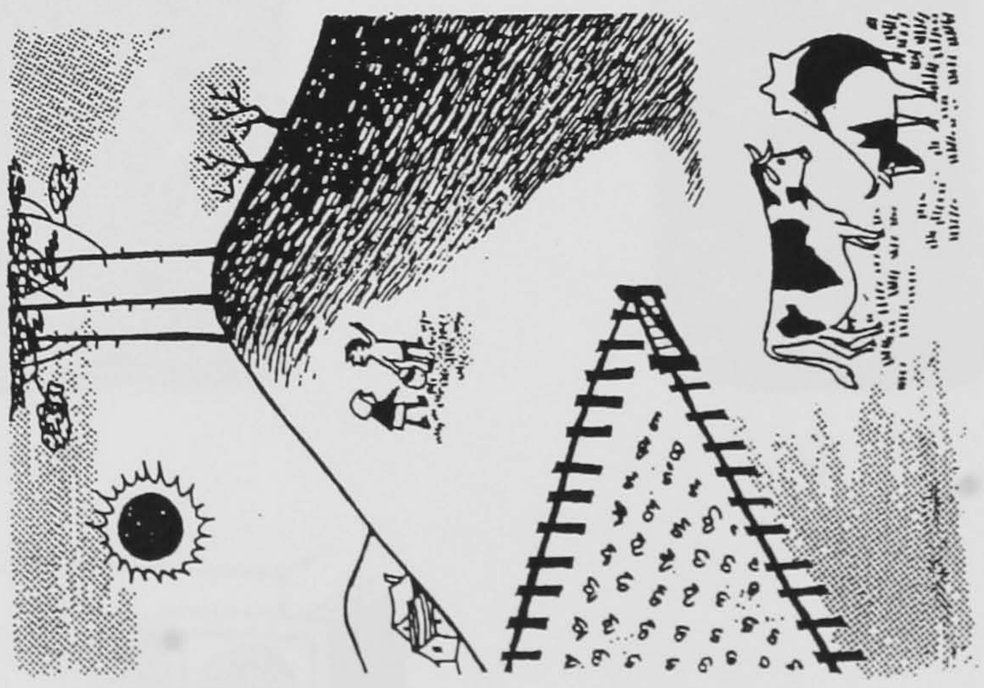

3


6


2

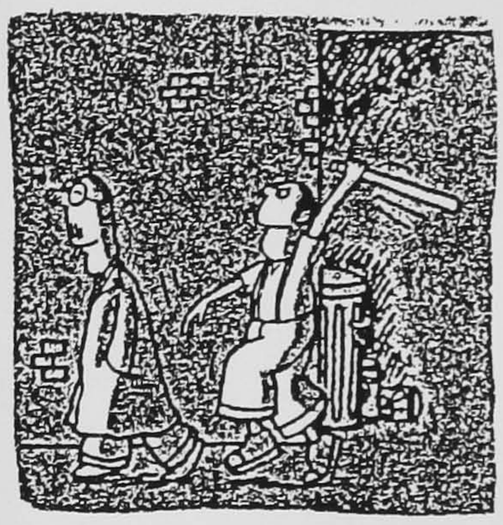

5


1


4

Unlucky Man

1



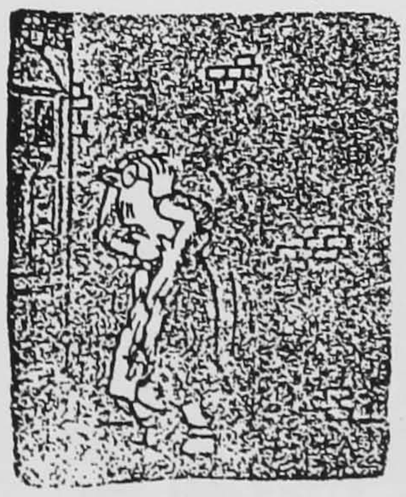
2

3

4

5

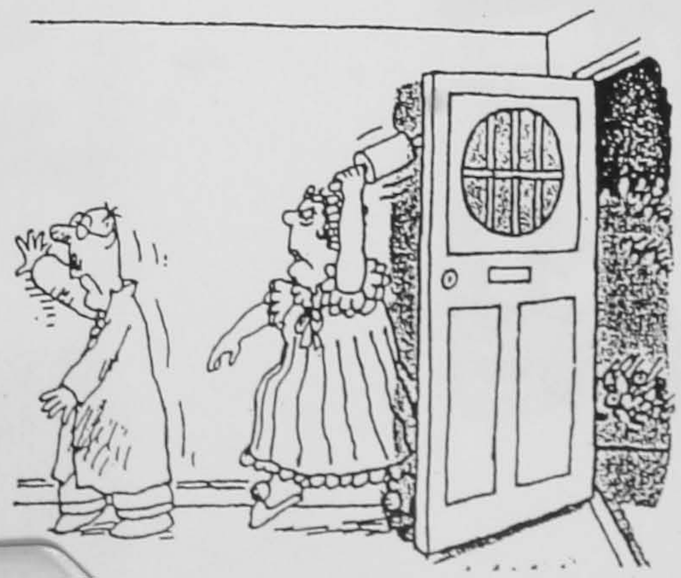
6



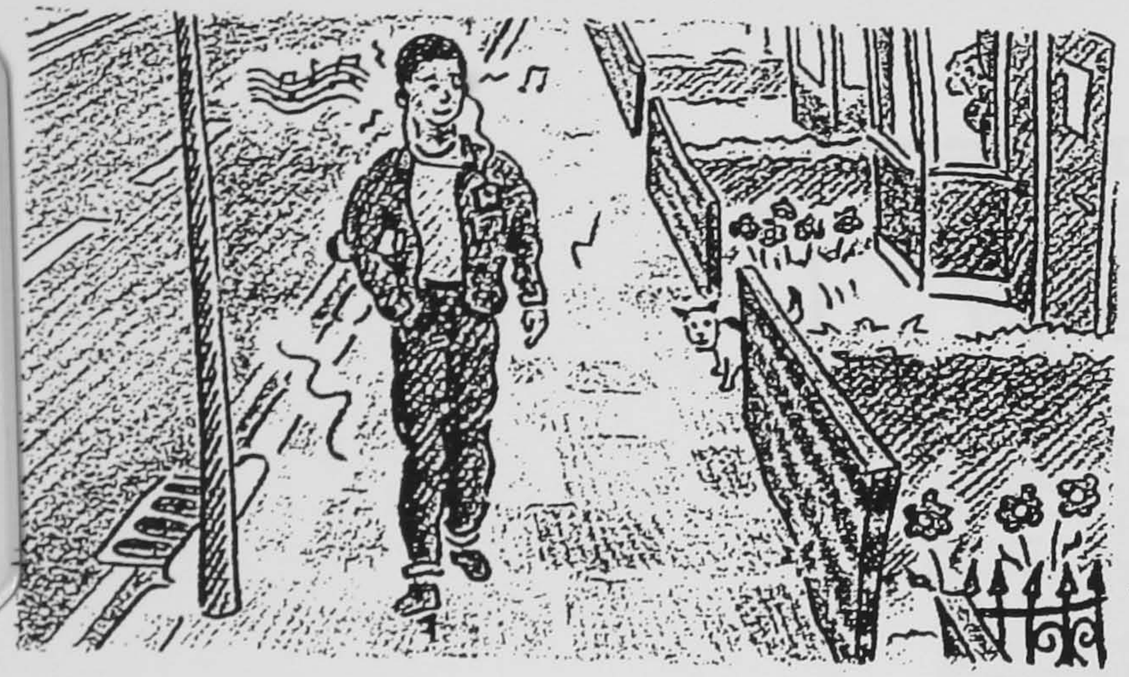
7

8

9



10

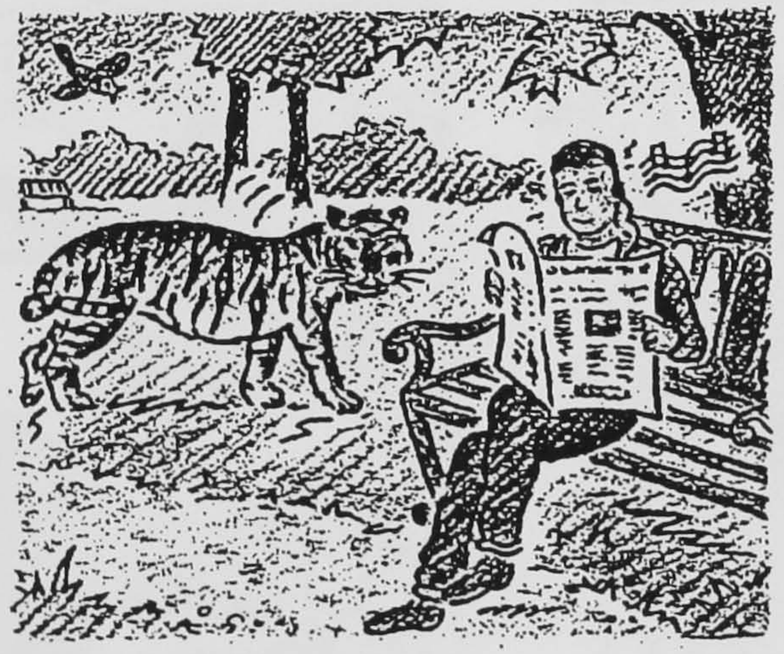


1

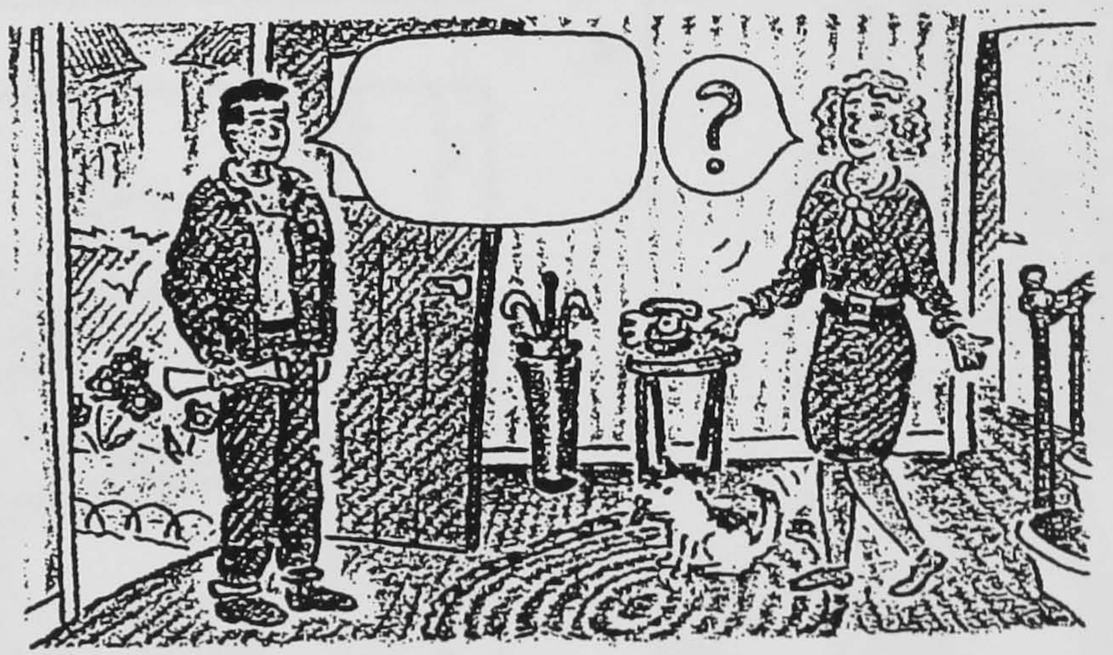
2

3

4



5



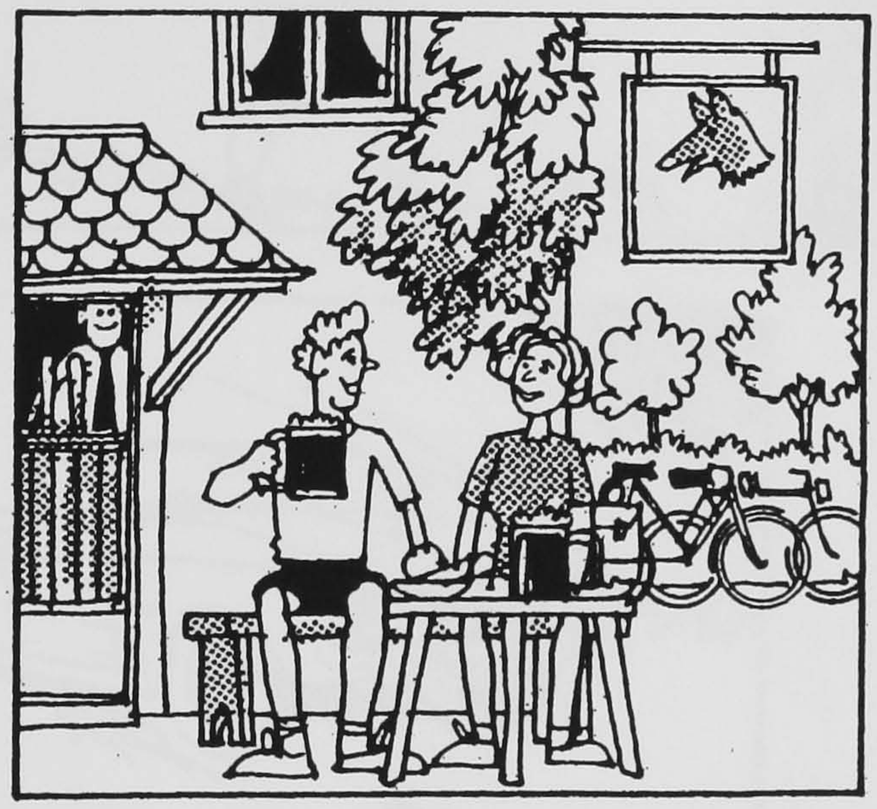
6

Walk man

Journey

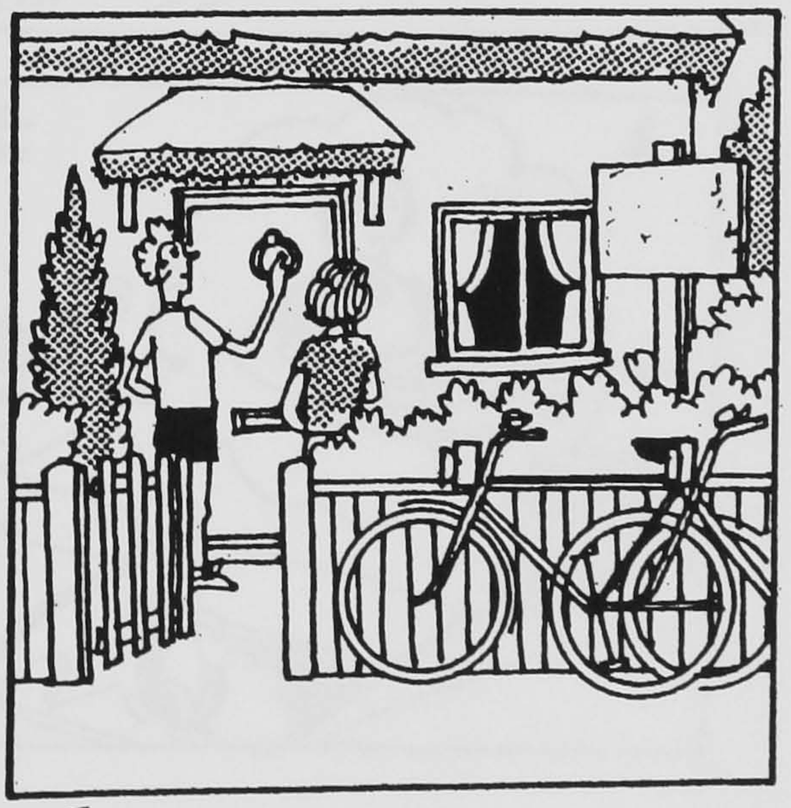
1

2



3

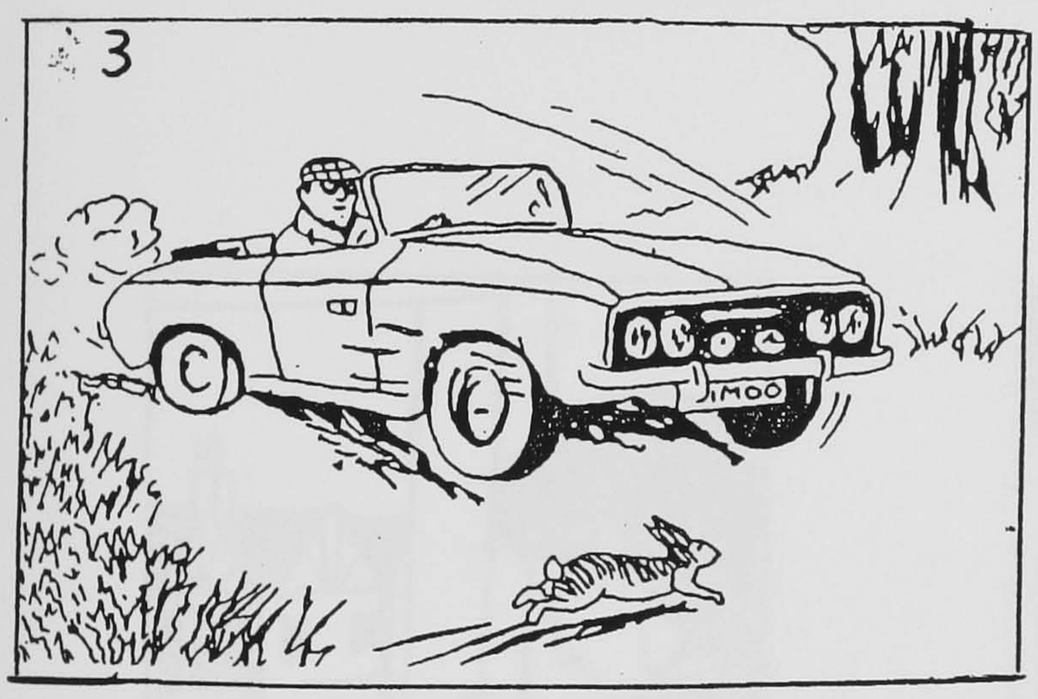
4

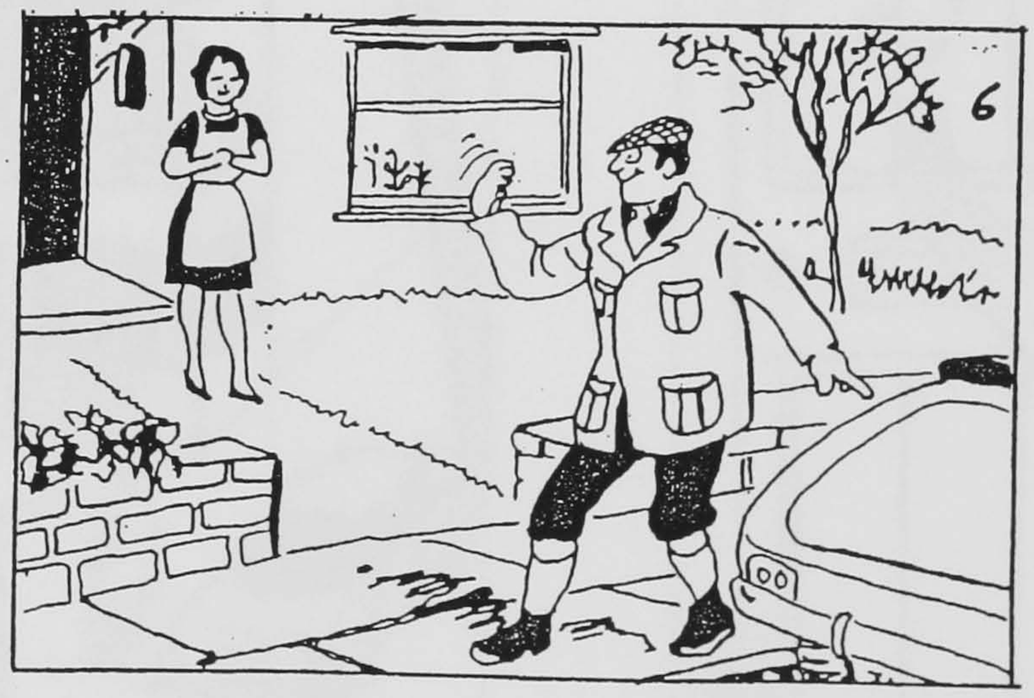


5

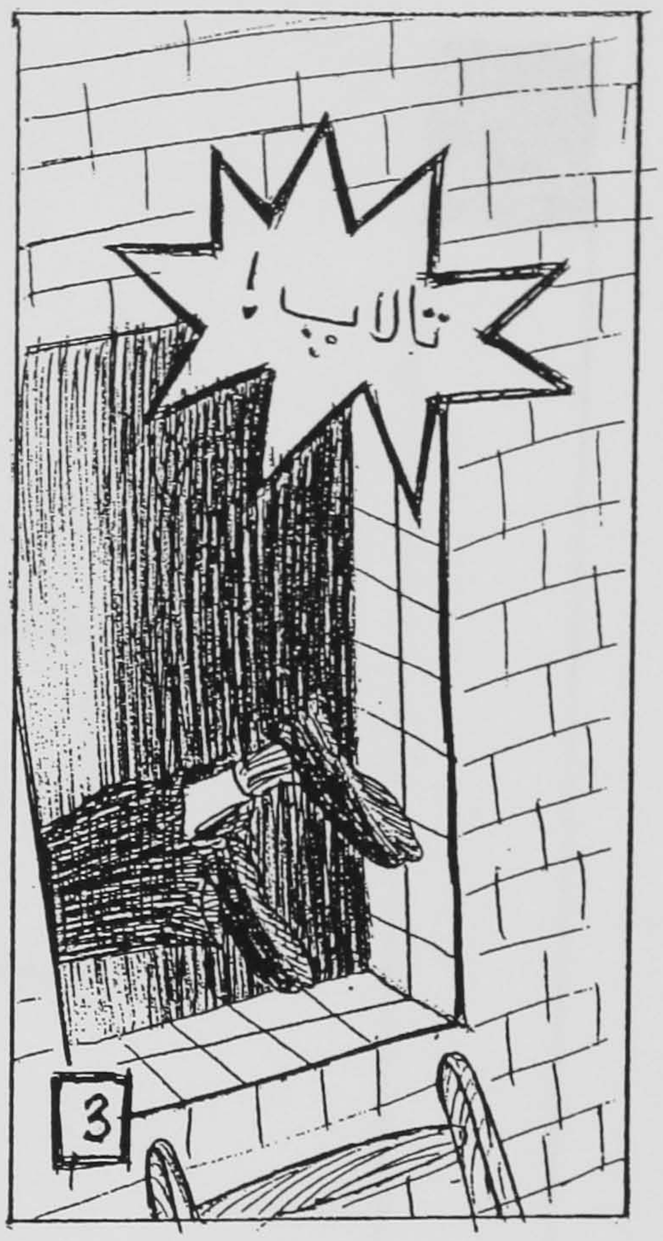
6

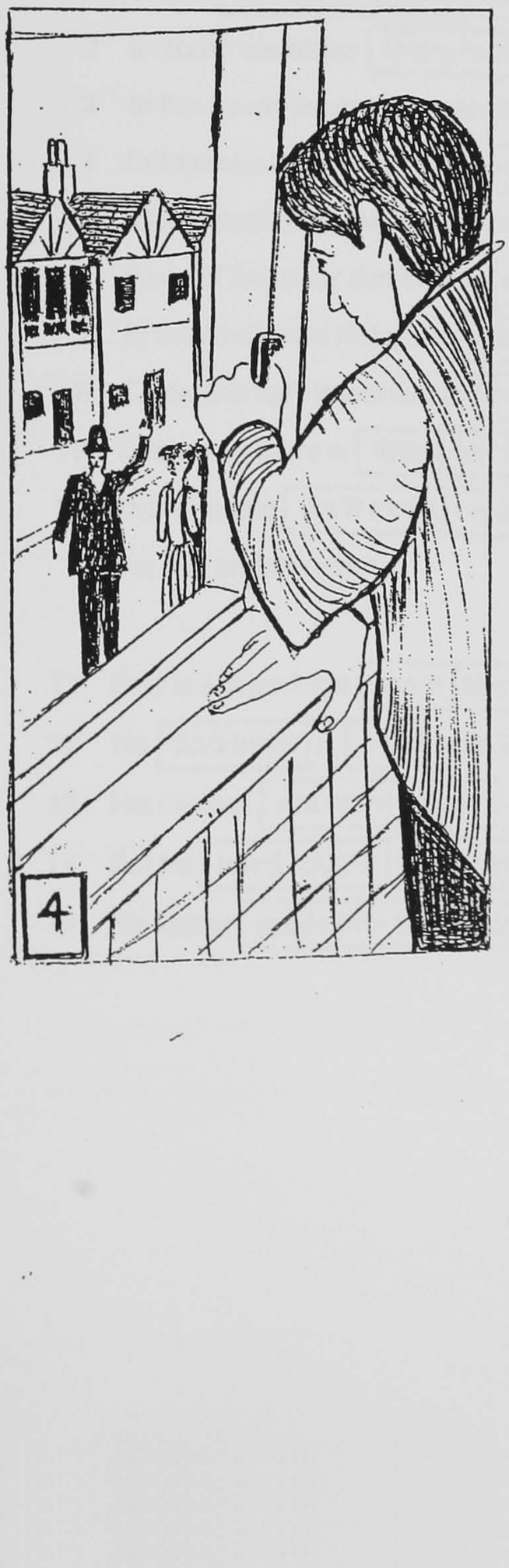
Hunting





Keys





Appendix 2

Oxford Placement Test 2

Grammar Test PART 1

Name _____

Total Listening	____ / 100	**Total Grammar Part 2**	____ / 50
Total Grammar Part 1	____ / 50	**Grand total**	____ / 200

Look at these examples. The correct answer is ticked.

a In warm climates people [like ✓] [likes] [are liking] sitting outside in the sun.

b If it is very hot, they sit [at] [in ✓] [under] the shade.

Now the test will begin. Tick the correct answers.

1 Water [be freezing] [is freezing] [freezes] at a temperature of 0°C. 1 ____

2 In some countries [there is] [is] [it is] dark all the time in winter. 2 ____

3 In hot countries people wear light clothes [for keeping] [to keep] [for to keep] cool. 3 ____

4 In Madeira they have [the good] [good] [a good] weather almost all year. 4 ____

5 Most Mediterranean countries are [more warm] [the more warm] [warmer] in October than in April. 5 ____

6 Parts of Australia don't have [the] [some] [any] rain for long periods. 6 ____

7 In the Arctic and Antarctic [it is] [there is] [it has] a lot of snow. 7 ____

8 Climate is very important in [most of] [most] [the most] people's lives. 8 ____

9 Even now there is [little] [few] [less] we can do to control the weather. 9 ____

10 In the future [we'll need] [we are needing] [we can need] to get a lot of power from the sun and the wind. 10 ____

11 Pele is still perhaps [most] [the most] [the more] famous footballer in the world. 11 ____

12 He [had been] [is] [was] born in 1940. 12 ____

13 His mother [not want] [wasn't wanting] [didn't want] him to be a footballer. 13 ____

14 But he [used] [ought] [has used] to watch his father play. 14 ____

15 His father [made him to] [made him] [would make him to] practise every day. 15 ____

subtotal ___/15

He learned to use [or his left foot or | and his left foot and | both his left foot and] his right. 16

He got the name Pele when he [had only ten years | was only ten | was only ten years]. 17

By 1956 he [has joined | joined | had joined] Santos and had scored in his first game. 18

In 1957 he [has been picked | was picked | was picking] for the Brazilian national team. 19

The World Cup Finals were in 1958 and Pele was looking forward [to play | to playing | to be playing]. 20

But he hurt [this | the | his] knee in a game in Brazil. 21

He thought he [isn't going to | couldn't | wasn't going to] be able to play in the finals in Sweden. 22

If he [hadn't been | weren't | wouldn't be] so important to the team, he would have been left behind. 23

But he was [a such | such a | a so] brilliant player, they took him anyway. 24

And [even though | even so | in spite of] he was injured, he helped Brazil to win the final. 25

The history of the World Cup is [quite a | a quite | quite] short one. 26

Football [has been | is being | was] played for 27

[above | over | more that] a hundred years, but the first World Cup 28

competition [did not be | was not | was not being] held until 29

1930. Uruguay [could win | were winning | had won] the Olympic football 30

final in 1924 and 1928 and wanted [be | being | to be] World Champions for the third time. 31

Four teams entered from Europe, but with [a little | few | little] success. 32

It was the first time [which | that | when] professional teams 33

[are playing | would play | had played] for a world title. 34

It wasn't until four years [later | more | further] that a 35

European team succeeded [to win | in winning | at winning] 36

for [the | a | its] first time. The 1934 World Cup was 37

again won by [a | the | one] home team, 38

[what | this | which] has been the case several times since 39

then. The 1934 final was [among | between | against] two 40

European teams, Czechoslovakia and Italy. Italy, [which | that | who] won, 41

went on [to win | winning | to have won] the 1938 final. Winning 42

successive finals is something that [is not | was not | has not been] achieved 43

again until Brazil did [these | them | it] in 1958 and 1962. If Brazil 44

[would have won | would win | had won] in 1966 then the 45

authorities would have needed to [have | let | make] the original World Cup replaced. 46

But England stopped the Brazilians [to get | getting | get] a third successive win. An England player, 47

Geoff Hurst, scored three goals in the final and won it almost [by his own | on himself | by himself]. 48

1966 proved [being | as being | to be] the last year that England 49

[would | will | did] even qualify for the finals till 1982, though they got in as winners in 1970. 50

subtotal ___/35

Grammar Test PART 2

51 Many [persons | people | peoples] nowadays believe that everyone should learn to use computers. 51

52 The majority of children in the UK [have | has | are having] access to a micro-computer. 52

53 There are more computers per head in England than [anywhere else | somewhere else | anywhere other] in the world. 53

54 Learning a computer language is not the same [as | like | than] learning a real language. 54

55 Most people start off with 'Basic', [who | what | which] is the easiest to learn. 55

56 Children seem to find computers easy, but many adults aren't used to [work | the work | working] with microtechnology. 56

57 There aren't [no | any | some] easy ways of learning how to program a computer. 57

58 The only way to become really proficient is to practise a lot [on your own | by your own | on your self]. 58

59 You can pick up the basics quite quickly if you [want to | would | are willing to] make an effort. 59

60 Most adults feel it would be easier if only they [would have started | would start | had started] computer studies earlier. 60

61 Some people would just [rather | prefer | better] not have anything to do with computers at all. 61

62 A lot have resigned themselves to never even [know | known | knowing] how a computer works. 62

63 Microtechnology is moving so fast that hardly [anybody | nobody | no one] can keep up with it all. 63

64 It's no use [in trying | to try | trying] to learn about computers just by reading books. 64

65 Everyone has [difficulty in learning | difficulties to learn | it difficult to learn] if they can't get 'hands-on' experience. 65

Below is a letter written to the 'advice' column of a daily newspaper. Tick the correct answers.

Dear Marge,

[I am writing | I will write | I should write] to you because I 66

[am not knowing | don't know | know not] what to do. I'm twenty-six and a teacher at 67

a primary school in Norwich where [I'm working | I've worked | I work] for the last five years. 68

When I [was | have been | had been] here for a couple of years, one of the older members of staff 69

[would leave | left | had been leaving], and a new teacher 70

[would be | became | was] appointed to work in the same department as me. 71

We [worked | have worked | should work] together with the same classes during her first year 72

and had the [opportunity for building | possibilities to build | chance to build] up a good professional 73

relationship. Then, about eighteen months after [she has arrived | to have arrived | arriving] 74

in Norwich, she decided to buy [her own | herself | her a] house. 75

subtotal /25

She was tired of [to live | live | living] in rented accommodation and wanted a place 76
[by her own | of her own | of herself]. At about the same time, I 77
[was given | have been given | gave] notice by the landlord of the flat 78
[what I was living | that I had lived | I was living] in 79
and she asked me if I [liked | had liked | would like] to live 80
with her. She [said | told | explained] me that by the time she 81
[would pay | would have paid | had paid] the mortgage 82
and the bills [it | there | they] wouldn't be 83
[a lot | many | few] left to live on. She suggested 84
[us to | we should | we may] share the house and share the costs. 85
It seemed like a good idea, so after [we'd agreed | we could agree | we agreed with] all the details 86
[what | that | who] needed to be sorted out, we moved into the new house together. 87
At the end of this month [we have lived | we have been living | we'll have been living] 88
together for a year and a half. It's the first time [I live | I'm living | I've lived] with anybody before, but 89
[I should guess | I might have guessed | I'd have guessed] what would happen. I've fallen in love with 90
her and now she's been offered another job 200 miles away and is going to move. I don't know what to
do. Please give me some advice.

Yours in shy desperation,

Steve

Look at the following examples of question tags in English. The correct form of the tag is ticked.

a He's getting the 9.15 train, [✓isn't he | hasn't he | wasn't he]?

b She works in a library, [isn't she | ✓doesn't she | doesn't he]?

c Tom didn't tell you, [hasn't he | didn't he | ✓did he]?

d Someone's forgotten to switch off the gas, [didn't one | didn't they | ✓haven't they]?

Now tick the correct question tag in the following 10 items:

91 Steve's off to China, [has he | hasn't he | isn't he]? 91

92 It'll be a year before we see him again, [won't it | won't we | shan't it]? 92

93 I believe he's given up smoking, [isn't he | don't I | hasn't he]? 93

94 I'm next on the list to go out there, [am not I | are I | aren't I]? 94

95 No doubt you'd rather he didn't stay abroad too long, [shouldn't you | wouldn't you | hadn't you]? 95

96 He's rarely been away for this long before, [is he | hasn't he | has he]? 96

97 So you think he'll be back before November, [shall he | will he | do you]? 97

98 Nobody's disagreed with the latest proposals, [did he | has he | have they]? 98

99 We'd better not delay reading this any longer, [should we | did we | had we]? 99

100 Now's hardly the time to tell me you didn't need a test at all, [did you | is it | isn't it]? 100

subtotal /25

Appendix 3

Samples of the Questionnaires

Questionnaire for the Unplanned Groups

Name: ……..
Age: ……..

Class Code: ……
Group Code: ……

Section One
Instructions:

Consider the four tasks of story telling you have just done. Please circle 1, 2, 3, or 4 to show what you think about them.

	very easy	easy	difficult	very difficult
A. Telling the story of the boys who play football is …..	1	2	3	4
B. Telling the story of the children who go on a picnic is ……	1	2	3	4
C. Telling the story of the man with the walkman is …..	1	2	3	4
D. Telling the story of the man who had an unlucky day is …	1	2	3	4

Section Two
Instruction:

Please write here any comments you have about telling each of the stories.

……

……

……

……

Questionnaire for the Planned Groups

Name: ……..
Age: ……..

Class Code: ……
Group Code: ……..

Section One

Instructions:

Consider the four tasks of story telling you have just done. Please circle 1, 2, 3, or 4 to show what you think about them.

	very easy	easy	difficult	very difficult
A. Telling the story of the boys who play football is …..	1	2	3	4
B. Telling the story of the children who go on a picnic is ……	1	2	3	4
C. Telling the story of the man with the walkman is …..	1	2	3	4
D. Telling the story of the man who had an unlucky day is …	1	2	3	4

Section Two

Instructions:

Please write here any comments you have about telling each of the stories.

……

……

……

……

Section Three

Instructions:

Read the following sentences and circle a, b, c, or d which best completes each sentence to explain your situation in telling the four stories.

The planning time I was given before telling the story of

	much better	somewhat better	a little better	not better at all
1. the boys who play football helped me tell it …	a	b	c	d
2. the children who go on a picnic helped me tell it …	a	b	c	d
3. the man who had an unlucky day helped me tell it …	a	b	c	d
4. the man who had a walkman helped me tell it ….	a	b	c	d

Appendix 4

Coding Symbols

AS units:	| |
Clause boundaries:	::
Pauses:	( )
Participant's experience:	[ ]
Researcher's experience:	{ }
False start:	#
Reformulation:	~
Replacement:	rpl
Hesitation:	^
Repetition:	*
Total time per task:	< >
Error-free clause:	errfr

Appendix 5

Samples of the Transcribed and Coded Data

Study One: Planned Group 3

Participant: S. DA.

Task 1: Unlucky Man

<33>

a man is walking in the street errfr | another man is er behind him errfr | (.72) he falls on the floor | he keeps on walking again errfr | (.4) another man er hit him with a er stick | (.68) he falls down again errfr | (.45) he reaches his home errfr | (.45) er his hand is on his head errfr | and he feels the pain errfr | (.5) he rings the door bell errfr | (.9) and he walks to his house errfr | (.68) his wife is er behind the door errfr :: waiting for him errfr | (.68) er and she is er going to hit him again errfr |

Task 2: Walkman

<40>

a man is walking in the street errfr | and er he is listening to his walkman errfr | (.54) two car crashes behind him in the street | (.72) er he's still listening to music errfr | (.77) there's a jewelry shop behind him in the street errfr | (.4) and someone er has stolen the er jewelry errfr | (.54) and there are policemen errfr :: (.63) fighting with the thieves errfr | (.5) and the man is still listening to his walkman errfr | (.86) now he is sitting on a bench in a park errfr | (.86) and er a tiger is behind him errfr | (1) er he reaches his house errfr | (.86) er and his wife (.68) er asking him some questions |

Task 3: Football

<42>

em there are five young boys errfr :: playing with a ball (.59) on a field errfr | suddenly their ball falls in to a hole (.63) er on the ground errfr | one of them is trying errfr :: to bring the ball out of the hole errfr | and the other two is (.4) looking at him | the fourth one is thinking er about errfr :: what to do errfr :: to bring it out of the hole errfr | er and he want to bring something | (.4) and the other one who is looking in the hole errfr :: er (.4) sees a snake errfr :: and he become frightened | (.4) the one who has er ran for help :: (.5) comes back with a bowl of water errfr :: and to pour it in the hole | he pours the water in the hole errfr | and the ball comes up errfr :: and they can er catch the ball errfr |

Task 4: Picnic

<52>

er there are two children in the picture errfr | they are preparing er jam and butter sandwiches errfr | (.59) er they're putting them in to the basket errfr :: to go to the picnic errfr | (.5) er their mom is er pouring hot tea in the flask for them errfr | (.4) while their mom is showing them the map er (.4) of er the place errfr :: they want to go to go * errfr :: (.68) their dog er goes in to the basket errfr | (.4) and they don't realised | (.4) er they say good bye to their mom errfr | (.4) and they start off :: to go to the picnic errfr | (.54) they climb up a hill errfr :: (.4) while is a sunny day | and two cows are on the hill errfr | (.4) they sit on the (1.1) floor # on the ground rpl errfr :: to have er something to eat errfr | (.77) and they see their dog errfr :: coming out of the basket errfr | (.54) when they look at it # (.45) to # at the ~ basket errfr :: (.63) to em bring the em food out of # out ~ :: (.4) they realize errfr :: that their dog (.45) er ate all their food |

Study Two: Foreground Tasks Group 2

Participant: G. IB.

Task 1: Hunting

<90>

I think errfr :: this story about # is about ~ a man errfr :: who wants errfr :: to hunt something for dinner errfr | you know errfr | he and his friends are going to hunting something a bird a rabbit anything for dinner | and his wife was waiting for him errfr| (.45) but when they went there errfr :: he wasn't able to just hunt anything |_(.47) because all the bords # birds rpl are fly | and he wasn't able to saw them | so he didn't catch anything for dinner errfr | and when he was driving back to home errfr :: and also # he was very sad | because he's know that :: his wife is now really angry errfr :: em to kill him | (.44) so em in the way to the bac- # em to the home rpl (.43) he saw a rabbit on the road | and he think | the rabbit's there errfr | so he just stop the car | took the rabbit | and put the rabbit # put it ~ at the back of er (.4) his car | and he said errfr | now we have really delicious (.41) dinner | and my wife will be really happy errfr | because it said | that good husband I have | (.45) em so he went back home errfr | and as he saw his wife errfr :: says ok | I have really delicious thing for dinner | I hunt a rabbit | (.4) and it's really hard I think errfr | none of my friends were able to hunt a rabbit errfr | and his wife was really happy and proud of him errfr | and said ok errfr | take it out of your car errfr | and as he opened the door of car :: the rabbit jumped down errfr | and then ran away errfr |

Task 2: Football

<69>

er I think errfr :: this story about two children :: who they were just playing em in front of their house and with a ball | and they were just kicking the ball errfr | and they try :: to catch it | but suddenly the ball just fall into a hole | and it was a deep hole errfr | they weren't able to catch it errfr | and to bring it up | so they were # they didn't ~ know errfr :: what to do errfr | because they are also frightened errfr :: to just put their er (.4) hand in that hole errfr | because they think errfr :: maybe there's a snake there errfr | or

something which dangerous | and because also deep | and they weren't able to bring it up | but one of the children was more cleverer than the other | he think about a very brilliant (.4) way :: to catch em the ball | and bring it up | he went errfr | and er (.42) bring a barrel of water # a bowl rpl full of water | and he said ok errfr | we can pour the water in the hole errfr | and so the ball will come up errfr | and because the hole will be full with the water the water * | and because the ball is very light errfr :: so it come up | and we can just catch it | so they did it errfr | and they were successful in that |

Task 3: Journey

<83>

this story is about a couple errfr :: who just wants :: to (.4) em spend their holiday in some way | they had one day off errfr | and they want errfr :: to go together to somewhere | to just be together er the day :: that they had errfr | (.48) so with his bicycles round sea | for example some places some interesting places errfr | and they # (.45) after that go swimming | and they had some plan for a # for their rpl day | they did so errfr | first they er ride their bicycle | and er # until they reach something a café on there # on the road rpl | and there they er drink a cup of er coffee or tea | and some sandwiches or something like that errfr | it was really good errfr | it was really delicious errfr | the sandwich was really good errfr | (.4) and after that they decide errfr :: to go for swimming | and so they went | and they really enjoyed their swimming errfr | it was very good errfr | the water was good errfr | the weather was good errfr | everything was just really perfect errfr | and after that they went to a house er errfr :: for just resting for some minute or for a night maybe | and they went to that house errfr | and there were just the old couple there errfr_| and they said ok welcome errfr | and they gave some breakfast to them | and they tell them :: what a # what an rpl interesting enjoyable day they had errfr | they were very surprised because of that errfr |

Appendix 6

Factor Analyses for Fluency Measures: Study Two

Factor Analysis for Journey

Rotated Component Matrix [a]

	Component		
	1	2	3
Jour/walk reformulat			-.883
Jour/Walk false start		.890	
Jour/walk replacement		.905	
Jour/walk repetition	.622		
Jour/walk length of run	-.617		.484
jour/walk speech rate	-.741		
Jour/walk no of pause mid clause	.824		
Jour/wak no of pause end clause	.660	.426	
Jour/walk total silence mid clause	.922		
Jour/walk total silence end clause	.855		
Jour/walk proper time speaking	-.853	.325	
Jour/walk pause length mid clause	.798		
Jour/walk pause length end clause	.603	-.303	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Factor Analysis for Walkman

Rotated Component Matrix [a]

	Component			
	1	2	3	4
Jour/walk reformulat		.847		
Jour/Walk false start	.311	.917		
Jour/walk replacement		.906		
Jour/walk repetition	.344	.583		-.479
Jour/walk length of run	-.665	-.363		
jour/walk speech rate	-.728	-.307		-.326
Jour/walk no of pause mid clause	.679	.504	.333	
Jour/wak no of pause end clause			.976	
Jour/walk total silence mid clause	.823	.406		
Jour/walk total silence end clause			.936	
Jour/walk proper time speaking	-.924			
Jour/walk pause length mid clause	.761			.377
Jour/walk pause length end clause				.877

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Factor Analysis for Hunting

Rotated Component Matrix[a]

	Component			
	1	2	3	4
HPREFORM				.951
HPREPETI		.701		
HPFALSTA			.807	.435
HPREPLAC			.922	
HPLOFRUN		-.804		
HPSPRATE	-.491	-.742		
HPNOFPS1	.667	.389		.498
HPNOFPS2	.792			
HPTOTSL1	.772	.486		.318
HPTOTSL2	.933			
HPTIMSPK	-.752	-.536		
HPPSLEN1	.573	.532	-.340	
HPPSLEN2	.792			

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 9 iterations.

Factor Analysis for Picnic

Rotated Component Matrix[a]

	Component		
	1	2	3
HPREFORM	.554		
HPREPETI	.777		
HPFALSTA	.872		
HPREPLAC	.808		
HPLOFRUN	-.648	-.415	
HPSPRATE	-.605	-.544	-.305
HPNOFPS1	.370	.491	.617
HPNOFPS2			.920
HPTOTSL1		.821	.466
HPTOTSL2			.942
HPTIMSPK		-.931	
HPPSLEN1		.704	
HPPSLEN2		.564	.647

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Factor Analysis for Football

Rotated Component Matrix[a]

	Component			
	1	2	3	4
FKREFORM				.773
FKFALSTA			.925	
FKREPLAC			.906	
FKREPETI	.317	-.476		
FKLOFRUN	-.665			.338
FKSPRATE	-.818			.330
FKNOFPS1	.780			.307
FKNOFPS2		.716		.425
FKTOTSL1	.910			
FKTOTSL2		.909		
FKPTMSPK	-.886			
FKPSLEN1	.689			
FKPSLEN2		.683		-.370

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

Factor Analysis for Keys

Rotated Component Matrix[a]

	Component			
	1	2	3	4
FKREFORM			.688	
FKFALSTA			.939	
FKREPLAC			.692	
FKREPETI	.705	-.338		
FKLOFRUN	-.800	-.370		
FKSPRATE	-.860			
FKNOFPS1	.824	.341		
FKNOFPS2		.765		
FKTOTSL1	.688			.657
FKTOTSL2		.948		
FKPTMSPK	-.602	-.396		-.596
FKPSLEN1				.919
FKPSLEN2		.740	-.344	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

Appendix 7

Correlations for Journey, Hunting and Football

Correlations

		corrected accuracy	COMPLEXI	VOCD	REFORMUL	FALSTART	SPEECHRA	LENOFRUN	TOTSIL1	TOTSIL2	NOPAUS1	NOFPAUS2
corrected accuracy	Pearson Correlation	1.000	-.081	.164	-.049	-.139	.222*	.379**	-.189	-.150	-.245*	-.061
	Sig. (2-tailed)	.	.450	.123	.643	.191	.036	.000	.075	.159	.020	.570
	N	90	90	90	90	90	90	90	90	90	90	90
COMPLEXI	Pearson Correlation	-.081	1.000	.217*	-.163	.177	.050	.011	-.134	-.129	-.060	-.101
	Sig. (2-tailed)	.450	.	.040	.124	.095	.639	.922	.208	.226	.577	.345
	N	90	90	90	90	90	90	90	90	90	90	90
VOCD	Pearson Correlation	.164	.217*	1.000	-.018	.261*	.146	.230*	-.093	.019	-.028	.139
	Sig. (2-tailed)	.123	.040	.	.867	.013	.170	.029	.385	.858	.795	.191
	N	90	90	90	90	90	90	90	90	90	90	90
REFORMUL	Pearson Correlation	-.049	-.163	-.018	1.000	.376**	-.010	-.040	.107	-.055	.252*	.043
	Sig. (2-tailed)	.643	.124	.867	.	.000	.929	.708	.317	.608	.017	.688
	N	90	90	90	90	90	90	90	90	90	90	90
FALSTART	Pearson Correlation	-.139	.177	.261*	.376**	1.000	.050	-.042	.155	.064	.356**	.098
	Sig. (2-tailed)	.191	.095	.013	.000	.	.638	.691	.143	.551	.001	.357
	N	90	90	90	90	90	90	90	90	90	90	90
SPEECHRA	Pearson Correlation	.222*	.050	.146	-.010	.050	1.000	.736**	-.631**	-.399**	-.557**	-.277**
	Sig. (2-tailed)	.036	.639	.170	.929	.638	.	.000	.000	.000	.000	.008
	N	90	90	90	90	90	90	90	90	90	90	90
LENOFRUN	Pearson Correlation	.379**	.011	.230*	-.040	-.042	.736**	1.000	-.395**	-.227*	-.410**	-.155
	Sig. (2-tailed)	.000	.922	.029	.708	.691	.000	.	.000	.031	.000	.144
	N	90	90	90	90	90	90	90	90	90	90	90
TOTSIL1	Pearson Correlation	-.189	-.134	-.093	.107	.155	-.631**	-.395**	1.000	.581**	.895**	.439**
	Sig. (2-tailed)	.075	.208	.385	.317	.143	.000	.000	.	.000	.000	.000
	N	90	90	90	90	90	90	90	90	90	90	90
TOTSIL2	Pearson Correlation	-.150	-.129	.019	-.055	.064	-.399**	-.227*	.581**	1.000	.532**	.760**
	Sig. (2-tailed)	.159	.226	.858	.608	.551	.000	.031	.000	.	.000	.000
	N	90	90	90	90	90	90	90	90	90	90	90
NOPAUS1	Pearson Correlation	-.245*	-.060	-.028	.252*	.356**	-.557**	-.410**	.895**	.532**	1.000	.463**
	Sig. (2-tailed)	.020	.577	.795	.017	.001	.000	.000	.000	.000	.	.000
	N	90	90	90	90	90	90	90	90	90	90	90
NOFPAUS2	Pearson Correlation	-.061	-.101	.139	.043	.098	-.277**	-.155	.439**	.760**	.463**	1.000
	Sig. (2-tailed)	.570	.345	.191	.688	.357	.008	.144	.000	.000	.000	.
	N	90	90	90	90	90	90	90	90	90	90	90

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations for Walkman, Picnic and Keys

Correlations

		corrected accuracy	COMPLEX	VOCD	REFORM	FALSTART	LENOFRUN	SPCHRATE	NOFPAUS1	NOFPAUS2	TOTSIL1	TOTSIL2	REPLACE	REPETITI	PAUSLEN1	PAUSLEN2
corrected accuracy	Pearson Correlation	1.000	.444**	-.083	-.099	-.310**	.376**	.409**	-.362**	-.172	-.382**	-.191	-.289**	-.302**	-.159	-.178
	Sig. (2-tailed)	.	.000	.435	.351	.003	.000	.000	.000	.106	.000	.071	.006	.004	.135	.093
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
COMPLEX	Pearson Correlation	.444**	1.000	-.090	-.047	-.229*	.207	.375**	-.317**	-.257*	-.378**	-.257*	-.160	-.265*	-.219*	-.133
	Sig. (2-tailed)	.000	.	.401	.663	.030	.051	.000	.002	.014	.000	.015	.133	.012	.038	.211
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
VOCD	Pearson Correlation	-.083	-.090	1.000	.082	.122	.151	.203	.088	.309**	-.056	.303**	.075	-.308**	-.123	.162
	Sig. (2-tailed)	.435	.401	.	.441	.251	.156	.055	.412	.003	.602	.004	.484	.003	.246	.127
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
REFORM	Pearson Correlation	-.099	-.047	.082	1.000	.713**	-.211*	-.091	.342**	.015	.230*	-.057	.371**	.389**	.002	-.056
	Sig. (2-tailed)	.351	.663	.441	.	.000	.046	.392	.001	.885	.029	.594	.000	.000	.982	.600
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
FALSTART	Pearson Correlation	-.310**	-.229*	.122	.713**	1.000	-.400**	-.369**	.589**	.186	.432**	.134	.817**	.483**	.061	.031
	Sig. (2-tailed)	.003	.030	.251	.000	.	.000	.000	.000	.079	.000	.209	.000	.000	.569	.773
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
LENOFRUN	Pearson Correlation	.376**	.207	.151	-.211*	-.400**	1.000	.739**	-.517**	-.285**	-.485**	-.302**	-.414**	-.416**	-.166	-.312**
	Sig. (2-tailed)	.000	.051	.156	.046	.000	.	.000	.000	.006	.000	.004	.000	.000	.119	.003
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
SPCHRATE	Pearson Correlation	.409**	.375**	.203	-.091	-.369**	.739**	1.000	-.577**	-.254*	-.644**	-.321**	-.345**	-.402**	-.283**	-.362**
	Sig. (2-tailed)	.000	.000	.055	.392	.000	.000	.	.000	.016	.000	.002	.001	.000	.007	.000
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
NOFPAUS1	Pearson Correlation	-.362**	-.317**	.088	.342**	.589**	-.517**	-.577**	1.000	.471**	.855**	.473**	.462**	.422**	.069	.282**
	Sig. (2-tailed)	.000	.002	.412	.001	.000	.000	.000	.	.000	.000	.000	.000	.000	.406	.007
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
NOFPAUS2	Pearson Correlation	-.172	-.257*	.309**	.015	.186	-.285**	-.254*	.471**	1.000	.422**	.901**	.183	.029	.127	.287**
	Sig. (2-tailed)	.106	.014	.003	.885	.079	.006	.016	.000	.	.000	.000	.084	.786	.231	.006
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
TOTSIL1	Pearson Correlation	-.382**	-.378**	-.056	.230*	.432**	-.485**	-.644**	.855**	.422**	1.000	.456**	.314**	.326**	.444**	.299**
	Sig. (2-tailed)	.000	.000	.602	.029	.000	.000	.000	.000	.000	.	.000	.003	.002	.000	.004
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
TOTSIL2	Pearson Correlation	-.191	-.257*	.303**	-.057	.134	-.302**	-.321**	.473**	.901**	.456**	1.000	.140	-.042	.180	.611**
	Sig. (2-tailed)	.071	.015	.004	.594	.209	.004	.002	.000	.000	.000	.	.188	.691	.090	.000
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
REPLACE	Pearson Correlation	-.289**	-.160	.075	.371**	.817**	-.414**	-.345**	.462**	.183	.314**	.140	1.000	.402**	.022	.052
	Sig. (2-tailed)	.006	.133	.484	.000	.000	.000	.001	.000	.084	.003	.188	.	.000	.838	.629
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
REPETITI	Pearson Correlation	-.302**	-.265*	-.308**	.389**	.483**	-.416**	-.402**	.422**	.029	.326**	-.042	.402**	1.000	-.009	-.113
	Sig. (2-tailed)	.004	.012	.003	.000	.000	.000	.000	.000	.786	.002	.691	.000	.	.936	.287
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
PAUSLEN1	Pearson Correlation	-.159	-.219*	-.123	.002	.061	-.166	-.283**	.089	.127	.444**	.180	.022	-.009	1.000	.211*
	Sig. (2-tailed)	.135	.038	.246	.982	.569	.119	.007	.406	.231	.000	.090	.838	.936	.	.046
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
PAUSLEN2	Pearson Correlation	-.178	-.133	.162	-.056	.031	-.312**	-.362**	.282**	.287**	.299**	.611**	.052	-.113	.211*	1.000
	Sig. (2-tailed)	.093	.211	.127	.600	.773	.003	.000	.007	.006	.004	.000	.629	.287	.046	.
	N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).